UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

Machine learning methods for quantitative structure-property relationship modeling

Ana Isabel Lino Teixeira

DOUTORAMENTO EM INFORMÁTICA
ESPECIALIDADE BIOINFORMÁTICA

2014

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

Machine learning methods for quantitative structure-property relationship modeling

Ana Isabel Lino Teixeira

DOUTORAMENTO EM INFORMÁTICA
ESPECIALIDADE BIOINFORMÁTICA

Tese orientada pelo Prof. Doutor André Osório e Cruz de Azerêdo Falcão e pelo Prof. Doutor João Paulo Arriegas Estevão Correia Leal.

2014

Resumo

Devido ao crescimento exponencial do número de compostos químicos descobertos diariamente e à morosidade/custo de medições experimentais, existe uma diferença significativa entre o número de compostos químicos conhecidos e a quantidade de compostos para os quais estão disponíveis propriedades experimentais. O desenvolvimento de novos métodos para a previsão de propriedades e organização de grandes coleções de moléculas que permitam revelar certas categorias/padrões químicos e selecionar amostras diversas/representativas para estudos exploratórios estão a tornar-se essenciais. Este trabalho tem como objetivo melhorar a capacidade de prever propriedades físicas, químicas e biológicas, através de métodos de aprendizagem automática aplicados a dados complexos não homogeneos (estruturas químicas), para grandes repositórios de informação.

Numa primeira fase deste trabalho, foi feito o estudo de metodologias atualmente aplicadas para a modelação quantitativa entre estrutura-propriedades. Estas metodologias tentam relacionar um conjunto seleccionado de descritores estruturais de uma molécula com as suas propriedades, utilizando uma abordagem baseada em modelos. Este trabalho centrou-se em solucionar as principais dificuldades identificadas na previsão de propriedades de compostos químicos e nas soluções exploradas utilizando diferentes representações moleculares, técnicas de seleção de descritores e abordagens de aprendizagem automática. Neste contexto, foi proposta uma abordagem híbrida inovadora para melhorar o capacidade de previsão e compreensão de problemas QSPR/QSAR utilizando o algoritmo *"Random Forests"* (Florestas Aleatórias) para seleção de descritores.

É reconhecido que, em geral, moléculas semelhantes tendem a ter propriedades semelhantes; assim, numa segunda fase deste trabalho foi

desenvolvida uma metodologia de aprendizagem automática baseada em instâncias para a previsão de propriedades de compostos químicos utilizando o espaço métrico construído a partir da semelhança estrutural entre moléculas. No entanto, este tipo de metodologia requer a quantificação de semelhança estrutural entre moléculas, o que é muitas vezes uma tarefa subjetiva, ambígua e dependente de julgamentos comparativos e, consequentemente, não existe atualmente nenhum padrão absoluto para definir semelhança molecular. Neste âmbito, foi desenvolvido um novo método de semelhança molecular, o *"Non-Contiguous Atom Matching Structural Similarity"* (NAMS), que se baseia no alinhamento de átomos utilizando algoritmos de emparelhamento que têm em conta os perfis topológicos das ligações e as características dos átomos e ligações. O espaço métrico molecular construído utilizando o NAMS pode ser aplicado à inferência de propriedades usando uma técnica de interpolação espacial, a *"krigagem"*, que tem em conta a relação espacial entre as instâncias, com o objetivo de se obter uma previsão consistente e interpretável, proporcionando uma melhor compreensão da relação entre estrutura-propriedades.

Palavras Chave: Aprendizagem Automática, Quimioinformática, Previsão de Propriedades, Relação Quantitativa entre Estrutura-Propriedade, Mineração de Dados Moleculares, Aprendizagem Baseada em Modelos, Aprendizagem Baseada em Instâncias, Semelhança Molecular, Sistemas de Informação

Abstract

Due to the high rate of new compounds discovered each day and the morosity/cost of experimental measurements there will always be a significant gap between the number of known chemical compounds and the amount of chemical compounds for which experimental properties are available. This research work is motivated by the fact that the development of new methods for predicting properties and organize huge collections of molecules to reveal certain chemical categories/patterns and select diverse/representative samples for exploratory experiments are becoming essential. This work aims to increase the capability to predict physical, chemical and biological properties, using data mining methods applied to complex non-homogeneous data (chemical structures), for large information repositories.

In the first phase of this work, current methodologies in quantitative structure-property modelling were studied. These methodologies attempt to relate a set of selected structure-derived features of a compound to its property using model-based learning. This work focused on solving major issues identified when predicting properties of chemical compounds and on the solutions explored using different molecular representations, feature selection techniques and data mining approaches. In this context, an innovative hybrid approach was proposed in order to improve the prediction power and comprehensibility of QSPR/QSAR problems using Random Forests for feature selection.

It is acknowledged that, in general, similar molecules tend to have similar properties; therefore, on the second phase of this work, an instance-based machine learning methodology for predicting properties of compounds using the similarity-based molecular space was

developed. However, this type of methodology requires the quantification of structural similarity between molecules, which is often subjective, ambiguous and relies upon comparative judgements, and consequently, there is currently no absolute standard of molecular similarity. In this context, a new similarity method was developed, the non-contiguous atom matching (NAMS), based on the optimal atom alignment using pairwise matching algorithms that take into account both topological profiles and atoms/bonds characteristics. NAMS can then be used for property inference over the molecular metric space using ordinary kriging in order to obtain robust and interpretable predictive results, providing a better understanding of the underlying relationship structure-property.

Keywords: Machine Learning, Cheminformatics, Molecular Data Mining, Property Prediction, Quantitative Structure-Property Relationship, Molecular Similarity, Model-Based Learning, Instance-Based Learning, Information Systems

Resumo Alargado

Devido ao crescimento exponencial do número de compostos químicos descobertos diariamente e à morosidade/custo de medições experimentais, existe uma diferença significativa entre o número de compostos químicos conhecidos e a quantidade de compostos para os quais estão disponíveis propriedades experimentais. O desenvolvimento de novos métodos para a previsão de propriedades e organização de grandes coleções de moléculas que permitam revelar certas categorias/padrões químicos e selecionar amostras diversas/representativas para estudos exploratórios estão a tornar-se essenciais. Este trabalho tem como objetivo melhorar a capacidade de prever propriedades físicas, químicas e biológicas, através de métodos de aprendizagem automática aplicados a dados complexos não homogeneos (estruturas químicas), para grandes repositórios de informação.

Numa primeira fase deste trabalho, foi feito um estudo das metodologias atualmente aplicadas para a modelação quantitativa entre estrutura e as suas propriedades (problemas QSPR/QSAR). Estas metodologias tentam relacionar um conjunto selecionado de descritores estruturais de um molécula com as suas propriedades físicas, químicas e biológicas, utilizando uma abordagem baseada em modelos. Para este efeito, existem três tarefas importantes que precisam ser tidas em conta: (1) a representação computacional de estruturas moleculares e a seleção de modelos de aprendizagem automática que consigam lidar com este tipo de dados; (2) a implementação e avaliação de modelos de previsão existentes utilizando diferentes casos de estudo com dados experimentais extraídos de diversas fontes e verificação da qualidade dos resultados produzidos; (3) implementação de ferramentas Web, que não só permitem o acesso aos dados experimentais de uma forma

compreensiva, mas também permitem estimar propriedades de compostos químicos, utilizando diferentes métodos de previsão sem a necessidade de conhecimento prévio sobre os detalhes de implementação desses métodos. Este trabalho centrou-se em solucionar as principais dificuldades identificadas na previsão de propriedades de compostos químicos e nas soluções exploradas usando diferentes representações moleculares, técnicas de seleção de descritores e abordagens de aprendizagem automática.

No âmbito deste trabalho foi proposto um método inovador para seleção de descritores baseado no algoritmo *"Random Forests"* que é utilizado para determinar a importância de cada variável no contexto de problemas QSPR/QSAR. Posteriormente, os modelos de previsão são treinados com Máquinas de Vectores de Suporte introduzindo sequencialmente as variáveis de acordo com a ordem pré-determinada baseada na importância de cada variável. Foram obtidos modelos preditivos robustos e interpretáveis utilizando um conjunto selecionado de descritores moléculares de tamanho muito inferior relativamente ao original, proporcionando uma melhor compreensão da relação subjacente entre a estrutura molecular e a propriedade em estudo. Para resolver este problema, foi desenvolvido um sistema de informação - ThermInfo (http://therminfo.lasige.di.fc.ul.pt) - que integra uma base de dados para propriedades estruturais e termoquímicas de compostos orgânicos e uma interface Web de fácil utilização. Este Sistema de Informação disponibiliza, a uma vasta comunidade, não só, acesso aos dados de uma forma organizada e compreensíva, mas também permite estimar propriedades utilizando diferentes métodos de previsão (por exemplo o "Extended Laidler Bond Additivity" (método ELBA)) sem a necessidade de conhecimento prévio sobre os métodos. O desenvolvimento de Sistemas de Informação em quimioinformática apresenta alguns desafios e requer a implementação de algumas funcionalidades que não estão presentes em outros sistemas de informação: (1) a representação da informação química,

(2) o armazenamento eficaz de estruturas químicas e de dados experimentais de uma forma compreensiva, (3) a recuperação eficaz de compostos químicos utilizando, por exemplo, o desenho de uma estrutura ou o seu nome tendo em conta que existem diversos sinónimos. A arquitetura e a tecnologia utilizada no desenvolvimento deste Sistema de Informação não são específicas para as propriedades termoquímicas e já foram adaptadas por outros investigadores para o armazenamento e recuperação de moléculas químicas e as suas propriedades relacionadas com penetração da barreira hemato-encefálica (B3Info: http://b3info.lasige.di.fc.ul.pt/), bem como a previsão deste tipo de propriedades (B3PP: http://b3pp.lasige.di.fc.ul.pt/).

A modelação *in silico* da permeação de moléculas através da barreira hemato-encefálica (BHE) é uma tarefa difícil, devido à complexidade do processo de permeação através da BHE e à informação incompleta e tendenciosa disponível. Vários estudos na literatura têm tentado prever a permeação através da BHE, no entanto com sucesso limitado e poucos, se alguns, com a aplicação prática a programas de descoberta e desenvolvimento de fármacos. Em parte, devido ao facto de que apenas cerca de 2% das moléculas existentes conseguirem atravessar a BHE e a maior probabilidade de moléculas mais pequenas o conseguirem, e aos conjuntos de dados disponíveis não representarem esta realidade. Estes conjuntos de dados são, geralmente, tendenciosos, uma vez que, sobre-representam as moléculas com a capacidade de permear a BHE. Para contornar esta limitação foi proposto um novo método baseado em estatística bayesiana, juntamente com métodos do estado da arte de aprendizagem automática para a produção de um modelo robusto capaz de ser aplicado em situações reais de procura de novos fármacos. Posteriormente, foi proposta uma importante extensão a esta metodologia, tentando determinar o número mínimo de descritores moleculares relevantes para a previsão da permeação de moléculas através da BHE. Para tal, foi utilizada a metodologia que determina a importância de cada variável com base no algoritmo *"Random Forests"*, e posteriormente foram treinados modelos utilizando

Máquinas de Vectores de Suporte introduzindo sequencialmente as variáveis de acordo com a ordem pré-determinada.

No âmbito deste trabalho participou-se num desafio do *8th Dialogue on Reverse Engineering Assessment and Methods* (DREAM 8) (junho-setembro de 2013), especificamente na tarefa NIEHS-NCATS-UNC Toxicogenética (https://www.synapse.org/#!Synapse:syn1761567). Este desafio tinha como objetivo modelar parâmetros de citotoxicidade populacional de fármacos utilizando descritores moleculares. Para esse efeito, foi aplicada a abordagem híbrida descrita acima de forma a melhorar a capacidade preditiva dos modelos reduzindo o número de descritores necessários (https://www.synapse.org/#!Synapse:syn2219104). Esta metodologia ficou classificada na primeira posição utilizando como critério de avaliação a raiz quadrada do erro quadrático médio e na segunda posição utilizando a raiz quadrada do erro quadrático médio e os coeficientes de correlação de Pearson e Spearman.

É reconhecido que, em geral, moléculas semelhantes tendem a ter propriedades semelhantes; assim, numa segunda fase deste trabalho foi desenvolvida uma metodologia de aprendizagem automática baseada em instâncias para a previsão de propriedades de compostos químicos utilizando o espaço métrico construído a partir da semelhança estrutural entre moléculas. No entanto, este tipo de metodologia requer a quantificação de semelhança estrutural entre moléculas, o que é muitas vezes uma tarefa subjetiva, ambígua e dependente de julgamentos comparativos e, consequentemente, não existe actualmente nenhum padrão absoluto para definir semelhança molecular. Neste âmbito, foi desenvolvido um novo método de semelhança molecular, o *"Non-Contiguous Atom Matching Structural Similarity"* (NAMS), que se baseia no alinhamento de átomos utilizando algoritmos de emparelhamento que têm em conta os perfis topológicos das ligações e as características dos átomos e ligações. O espaço métrico molecular

construído utilizando o NAMS pode ser aplicado à inferência de propriedades usando uma técnica de interpolação espacial, a *"krigagem"*, que tem em conta a relação espacial entre as instâncias com o objectivo de se obter uma previsão consistente e interpretável, proporcionando uma melhor compreensão da relação entre estrutura-propriedades. Os resultados globais de previsão obtidos com esta metodologia, utilizando diferentes casos de estudo, encontram-se dentro das margens de confiança de estudos QSPR/QSAR encontrados na literatura. No entanto, a abordagem apresentada mostrou várias vantagens relativamente às abordagens QSPR/QSAR habituais, nomeadamente: (1) faz uso da semelhança ou distância entre os compostos e não é necessário utilizar qualquer método de seleção de descritores ou ter qualquer conhecimento prévio do problema ou propriedade a prever; por isso, pode ser diretamente aplicado à maioria dos estudos QSPR/QSAR e a qualquer composto (mesmo que nunca tenham sido sintetizados), desde que a sua fórmula estrutural seja conhecida; (2) o mapa de semelhanças que posiciona cada molécula do conjunto de dados no espaço métrico pode ser usado para prever qualquer propriedade física, química ou biológica de compostos, desde que estejam disponíveis os dados experimentais dessas mesmas propriedades; (3) é possível identificar as situações em que os erros de previsão são considerados elevados (medida de extrapolação) estimando a variância da *"krigagem"* para cada previsão - uma variância estimada elevada indica que o composto se encontra fora do domínio de aplicabilidade do modelo, uma vez que não existem compostos semelhantes ou os compostos mais semelhantes têm propriedades muito variáveis, enquanto uma variância estimada baixa indica que o modelo é capaz de prever o valor da propriedade com elevada confiança; (4) o modelo é facilmente compreensível relativamente a um modelo de *caixa preta*, uma vez que é possível identificar os compostos que mais contribuiram para a previsão; (5) novos compostos podem ser facilmente incluídos ou removidos ao conjunto de treino; nesta abordagem a função alvo é aproximada localmente para cada composto de teste, em vez de gerar

um modelo global que precisa ser re-treinado sempre que o conjunto de treino é alterado; (6) o método pode ser aplicado a conjuntos de dados de qualquer dimensão, no entanto os resultados preditivos têm maior probabilidade de melhorar com o aumento do número de instâncias de treino uma vez que, a probabilidade de encontrar compostos mais semelhantes também aumenta; (7) a procura de relações entre a estrutura e propriedade é realizada num espaço de hipóteses mais rico, em vez de aproximar uma função altamente parametrizada num espaço de hipóteses único. Esta abordagem pode, simultaneamente, resolver diferentes problemas e lidar com sucesso com as mudanças no domínio do problema.

Em suma, as principais contribuições deste trabalho são: (1) uma abordagem inovadora para melhorar o capacidade de previsão e compreensão de problemas QSPR/QSAR utilizando o algoritmo *"Random Forests"* para seleção de descritores; (2) o desenvolvimento de um Sistema de Informação (ThermInfo) para coligir, recuperar, e prever dados termoquímicos; (3) o desenvolvimento de um método inovador para calcular semelhança estrutural (NAMS) com base no alinhamento entre átomos de moléculas e uma ferramenta Web que disponibiliza esta metodologia para a comunidade; (4) um novo método para prever propriedades físicas, químicas ou biológicas de moléculas utilizando o espaço métrico construído a partir da semelhança estrutural entre moléculas; (5) a colaboração no desenvolvimento de novas abordagens com base em estatística bayesiana, juntamente com métodos de aprendizagem automática e seleção de descritores para a produção de modelos robustos em cenários reais de pesquisa de novos fármacos; (6) a colaboração no desenvolvimento de um Sistema de Informação para disponibilizar e prever dados de penetração de compostos na barreira hemato-encefálica (B3Info).

O trabalho aqui apresentado consiste no primeiro passo para conceber um novo sistema de desenvolvimento e procura novas moléculas com propriedades químicas, físicas ou biológicas alvo, permitindo testar

milhares de pequenas alterações à estrutura de compostos com um custo mínimo. Esta abordagem também permite a compreensão da relação entre a estrutura e propriedades contribuíndo para o avanço do conhecimento sobre a propriedade em análise e o mecanismo de ação do composto.

Acknowledgements

In this doctoral dissertation, I present results from four years of my research into machine learning methods to quantitative structure-property relationship modelling. Most of all, it was a great time, and I hope to convey to you, the reader, part of the excitement, curiosity, and satisfaction that I experienced during this time. In recognition of my PhD advisors' contributions to this work this document is written in the first person plural. A doctoral project is a major undertaking and during these years, many people contributed directly with helpful input, productive cooperation, constructive criticism or indirectly with friendship and moral support and got to deserve my gratitude.

Undoubtedly, first of all I want to thank my main advisor, Dr. André Falcão. I was lucky to have a great visionary in all matters of machine learning and a great scientist with an inexhaustible capacity to conceive new interesting ideas to solve all kind of problems. Dr. André is confident, pragmatic, rigorous, honest, enthusiastic, and he seems to be able to solve any riddle. He was always available and willing to discuss any ideas and issues and he was always very supportive when things did not go as hoped for. He taught me how science works, how to write papers, how to communicate science, how to see bad results as opportunities. Dr. André always gave me the freedom to pursue interesting avenues of inquiry but also keeping me on track. The freedom to follow up on different ideas of my and his own has made the last four years more fun than work and I feel that I could continue working on this project for a long time without getting tired. His advice during these years has helped me out of seemingly dead ends and I have certainly learnt a lot from our conversations. It has

been a pleasure working with Dr. André and I look forward to fruitful collaborations in the near future.

I would also like to thank my co-advisor Dr. João Paulo Leal as he has also played an important role in these last years. Dr. João Paulo closely collaborated on the chemical aspects of the research and supported me with valuable advice, encouragement and suggestions.

I am also grateful to Dr. Francisco Couto whom undeniably helped me to realize how important and challenging is applying informatics to the study of complex biological and chemical systems when I was a Biology graduate student. Furthermore, he advised me during my master's thesis work and introduced me to the cheminformatics research line. He also motivated me to pursue a PhD and advised me to work with Dr. André Falcão. Thank you for everything!

I am very grateful to the master students that contributed directly to this work and allowed me to teach and learn at the same time: Inês Martins, Luís Pinheiro and Rony Reis.

I would like extend my thanks to the Molecular Energetics group of Centro de Química e Bioquímica, particularly to Dr. Rui Centeno Santos, Prof. Dr. José Artur Martinho Simões and Prof. Dr. João Paulo Leal for the thermochemical properties dataset and for conceiving the idea of developing an information system to collect and retrieve thermochemical data (ThermInfo) and choosing LaSIGE, particularly me, to collaborate in its implementation. This collaboration introduced me to the cheminformatics research line and I am very grateful for that.

I would also like to thank my thesis committee, Dr. João Aires de Sousa, Dr. Ana Paula Afonso and Dr. Graça Gaspar for their valuable feedback and suggestions during my qualification step.

I also gratefully acknowledge Dr. Yvonne Martin, Dr. Robert Brown, and Abbott for the monoamine oxidase inhibitor dataset.

Many colleagues and ex-colleagues contributed to the good atmosphere in LaSIGE and for that I am very gratefully. I am particularly grateful to those who contributed more closely to a diverse panoplia of different activities ranging from work or personal related conversations and discussions, meetings and tertulias, coffee- and lunch-breaks, among others: Bruno Tavares, Carlos Costa, Cátia Machado, Cátia Pesquita, Daniel Faria, Hugo Bastos, Jeferson Souza, João Ferreira, Juliana Duque, Lélio Duane, Luís Filipe Lopes, Nuno Silva, Patrícia Sousa, Rony Reis and Tiago Grego.

I would like extend my thanks to Pedro Gonçalves for the efficiency demonstrated in solving administrative and bureaucratic issues, as well as to ADMIN DI-FCUL for all the help with server configurations.

I want to thank LaSIGE for the opportunity to integrate the XLDB group, for giving me the possibility to work in the interesting field of modelling structure-property relationship using machine learning methods, and providing me with financial support for publications, facilities and training needed for my research.

I would also like to thank Centro de Química e Bioquímica for financial support to purchase a dedicated server for this research work.

I am also grateful to the Faculdade de Ciências da Universidade de Lisboa which has been my second home since I finished high school and all the teachers who contributed to my academic education.

I gratefully acknowledge the Fundação para a Ciência e Tecnologia for funding my PhD (doctoral grant SFRH/BD/64487/2009), without which my research would not have been possible.

Finally, I would like to express my gratitude to Emir Catak, my friends and family for their support, encouragement and inspiration. In particular to my dear parents, Maria José and José Alberto, for raising me with love, affection, understanding, character and values, as well as all the dedication and encouragement to reach higher which was imperative for me to get here. Also, to my dearest brother, José

Miguel, for his support, companionship, humour, encouragement and willingness to help, as well as for all relaxing moments. It is to them I dedicate this completed dissertation. A special dedication goes also to my aunt Maria Silvanira (R.I.P., 27-07-2014) who fought a cancer like a warrior but could not wait any longer "to see my future after this PhD", as she told me many times...

"Eu não sei o que é que os outros pensarão lendo isto. Mas acho que isto deve estar bem porque o penso sem estorvo, nem ideia de outras pessoas a ouvir-me pensar. Porque o penso sem pensamentos, porque o digo como as minhas palavras o dizem."

$\sim$ Alberto Caeiro (Heterónimo de Fernando Pessoa), in "Poemas Inconjuntos"

Ana Isabel Lino Teixeira

Lisbon, 30^{th} July of 2014

Neither a lofty degree of intelligence nor imagination nor both together go to the making of genius.

Love, love, love, that is the soul of genius.

$\sim$ Wolfgang Amadeus Mozart

Contents

1 Introduction **1**

1.1 Context: Data Explosion in Chemistry 1
1.2 Motivation 3
1.3 Problem Statement and Aims of the Study 5
1.4 Methodology Overview 9
1.5 Contributions and Publications 12
1.6 Overview of the Document 16

2 Background **19**

2.1 Quantitative Structure-Property/Activity Relationship Methods . 20
2.2 Approaches to represent chemical structures 24
2.2.1 Generation of computer molecular representations and descriptors from structure 26
2.2.1.1 Molecular Graph 26
2.2.1.2 Graph Tables and Matrices 30
2.2.1.3 Molecular Descriptors 33
2.2.1.4 Molecular Fragments 36
2.2.2 Molecular Similarity 39
2.2.2.1 Similarity according to constitution 40
2.2.2.2 Similarity according to configuration and conformation 40
2.2.2.3 Similarity Metrics 41
2.3 Approaches to select relevant molecular descriptors 44
2.3.1 Dimensionality Reduction Methods 45
2.3.1.1 Principal Component Analysis 45

2.3.1.2 Partial Least Squares . . . 45
2.3.1.3 Multidimensional Scaling . . . 46
2.3.1.4 Self-Organizing Map . . . 46
2.3.2 Feature Selection Methods . . . 47
2.3.2.1 Correlation- or Covariance-Based Methods . . . 47
2.3.2.2 Genetic Algorithms . . . 47
2.3.2.3 Simulated Annealing . . . 48
2.4 Approaches to establish structure-property relationships using multivariate methods . . . 49
2.4.1 Instance-based Learning Approaches . . . 50
2.4.1.1 k-Nearest Neighbours . . . 50
2.4.2 Model-based Learning Approaches . . . 51
2.4.2.1 Multiple Linear Regression . . . 51
2.4.2.2 Partial Least Squares . . . 51
2.4.2.3 Artificial Neural Networks . . . 52
2.4.2.4 Support Vector Machines . . . 52
2.4.2.5 Random Forests . . . 53
2.5 Approaches to assess and validate QSPR/ QSAR models . . . 54
2.5.1 Model fit . . . 55
2.5.1.1 Regression models . . . 55
2.5.1.2 Classification models . . . 56
2.5.2 Internal cross-validation . . . 58
2.5.3 External Validation . . . 58
2.5.4 Y-randomization . . . 59
2.5.5 Applicability Domain . . . 59
2.6 Summary . . . 60

3 Model-based Methods for Quantitative Structure-Property Relationship Modelling 61
3.1 Related Work in QSPR/QSAR Modelling . . . 62
3.2 Feature Selection using variable importance . . . 65
3.3 Data and Methods . . . 67
3.3.1 Data and data pre-processing . . . 68

3.3.2 Modelling Methodology 69
3.3.2.1 Approaches for model generation 69
3.3.2.2 Approaches for feature selection 72
3.3.2.3 Model validation and evaluation 74
3.4 Results 75
3.4.1 Case A1 - Predicting Enthalpy of formation for Hydrocarbon compounds 76
3.4.1.1 Model development without a feature selection or dimensionality reduction step 76
3.4.1.2 Model development with a feature selection or dimensionality reduction step 76
3.4.1.3 Model Validation with an Independent Validation Set 82
3.4.2 Case A2 - Predicting Enthalpy of formation and phase change for ThermInfo's dataset 83
3.4.2.1 Model development without a feature selection or dimensionality reduction step 83
3.4.2.2 Model development with a feature selection step . 83
3.4.2.3 Model Validation with an Independent Validation Set 86
3.4.3 Case D - NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge 87
3.4.3.1 Model development without a feature selection or dimensionality reduction step 87
3.4.3.2 Model development with a feature selection or dimensionality reduction step 88
3.4.3.3 Model Validation with an Independent Validation Set 91
3.4.4 Case G - Blood-Brain Barrier (BBB) Penetration Modeling 92
3.4.5 Web-based Information Systems and Tools 97
3.4.5.1 ThermInfo: Collecting, Retrieving, and Estimating Reliable Thermochemical Data 97

3.4.5.2 B3Info – An information system for molecular Blood-Brain Barrier penetration data 100
3.5 Discussion 101
3.5.1 Case A1 - Predicting Enthalpy of formation for Hydrocarbon compounds 101
3.5.1.1 Selected chemical descriptors 101
3.5.1.2 Prediction errors analysis 103
3.5.2 Case A2 - Predicting Enthalpy of formation and phase change for ThermInfo's dataset 105
3.5.2.1 Selected chemical descriptors 105
3.5.2.2 Prediction errors analysis 106
3.5.3 Case D - NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge 108
3.5.4 Case G - Blood-Brain Barrier (BBB) Penetration Modelling 110
3.6 Summary 111

4 Representing the Molecular Space Based on Structural Similarity 113
4.1 Measuring Molecular Similarity 114
4.1.1 Approaches based on structural descriptors 115
4.1.2 Approaches based on molecular fragments 115
4.1.3 Approaches based on graph matching 117
4.1.4 Quantifying the degree of similarity/dissimilarity between molecules 118
4.1.5 Non-contiguous atom matching structural similarity 119
4.2 Development of the Non-contiguous Atom Matching Structural Similarity 121
4.2.1 Concepts and overview 121
4.2.2 Molecular Alignment by Bond Matching 123
4.2.2.1 Bond Similarity 124
4.2.2.2 *Aba-bond* distance-compensation functions 124
4.2.2.3 *Aba-Bond* Matching Function 125
4.2.3 Molecular Alignment by Atom Matching 126

4.2.4 Translating the Molecular Alignment into a Structural Similarity Score . . . 127
4.2.5 An illustrative example of the application of NAMS . . . 128
4.3 Implementation . . . 133
4.3.1 Implementation of Fingerprints-based Structural Similarity 134
4.3.2 Implementation of Non-contiguous Atom Matching Structural Similarity . . . 135
4.3.2.1 Computational efficiency . . . 140
4.4 Results and Discussion . . . 140
4.4.1 Case-Study A1 - Discriminate molecules with repeated substructures . . . 141
4.4.2 Case-study E - Discriminate similar molecules with different activity levels . . . 144
4.4.3 Case-study F - Molecular similarity for inference . . . 146
4.4.4 Web Tool . . . 148
4.5 Summary . . . 149

5 Instance-based Methods for Quantitative Structure-Property Relationship Modelling 153
5.1 From Similarity to Property Prediction . . . 155
5.2 Methods . . . 158
5.2.1 Modelling Methodology . . . 158
5.2.1.1 Ordinary Kriging . . . 159
5.2.1.2 Semivariogram . . . 163
5.2.1.3 Neighbourhood Selection Strategies . . . 164
5.2.2 Implementation of Ordinary Kriging . . . 166
5.2.2.1 *CoordKrig* - Coordinate based kriging . . . 166
5.2.2.2 *DistKrig* - Distance based kriging . . . 167
5.2.3 Molecular Representation . . . 167
5.2.3.1 A. Structural similarity based on molecular descriptors . . . 168
5.2.3.2 B. Structural similarity based on molecular fragments . . . 168

5.2.3.3 C. Structural similarity based on graph matching 169
5.2.4 Model Validation 169
5.3 Data 170
5.4 Results 170
5.4.1 Case A1 - Predicting enthalpy of formation of gas phase . 170
5.4.2 Case B - Predicting Aqueous Solubility 173
5.4.3 Case C - Predicting Dihydrofolate reductase (DHFR) inhibitors activity 177
5.5 Discussion 182
5.5.1 Neighbourhood selection strategies 184
5.5.2 Prediction error analysis 186
5.5.3 Relationship between prediction error and molecular similarity 189
5.5.4 Kriging estimated variance and its relationship with prediction errors 192
5.5.5 Effect of the training set size on the predictive results . . . 195
5.5.6 Assumptions and Limitations 196
5.6 Summary 199

6 Conclusions **203**
6.1 Overall Approach 203
6.2 Summary of Research Contributions 213
6.3 Limitations and Future Work 213

A Case-Studies **217**
A.1 Case-studies, Data and Data pre-processing 217
A.1.1 Case A - Predicting Thermochemical Properties 218
A.1.1.1 Case A1 - Predicting Enthalpy of formation for Hydrocarbon compounds 221
A.1.1.2 Case A2 - Predicting Enthalpy of formation and phase change for ThermInfo's dataset 224
A.1.2 Case B - Predicting Aqueous Solubility 226
A.1.3 Case C - Predicting Dihydrofolate Reductase (DHFR) Inhibition Activity 228

A.1.4 Case D - NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge . . . 230
A.1.5 Case E - Steroids and their binding affinity to the corticosteroid binding globulin (CBG) receptor . . . 234
A.1.6 Case F - Classification of Monoamine Oxidase (MAO) Inhibition Level based on Molecular Similarity . . . 236
A.1.7 Case G - Blood-Brain Barrier (BBB) Penetration Modelling 237
A.2 Molecular descriptors: implementation and pre-processing . . . 239
A.2.1 Descriptor set A - Molecular descriptors calculated by E-DRAGON . . . 240
A.2.2 Descriptor set B - Structural descriptors calculated by Openbabel . . . 241
A.2.3 Descriptor set C - Molecular descriptors calculated by Chemistry Development Kit (CDK) . . . 241
A.2.4 Descriptor set D - Daylight Fingerprints . . . 241
A.2.5 Descriptor set E - Simplex representation of molecular structure (SiRMS) . . . 242
A.2.6 Descriptor set F - Extended Laidler Bond Additivity (ELBA) parameters . . . 242
A.2.7 Descriptor set G - Chemical Space Mapping based on Similarity/Dissimilarity . . . 244

B Datasets and Implementation Details 245
B.1 Case-study A1: Datasets . . . 245
B.2 Case-study A2: Datasets . . . 246
B.3 Case-study B: Datasets . . . 246
B.4 Case-study C: Datasets . . . 247
B.5 Case-study E: Datasets . . . 247
B.6 Case-study G: Datasets . . . 247
B.7 Molecular descriptors F - Extended Laidler Bond Additivity (ELBA) parameters . . . 248
B.8 NAMS - Atom Substitution Matrices (ASM) . . . 248

C Results Details **249**

C.1 Case-study A1: List of selected descriptors in model-based learning approach 249

C.2 Case-study A2: Detailed results for model-based learning approach 249

C.3 Case-study G: Detailed results for model-based learning approach 250

C.4 Case-study A1: Detailed results for neighbourhood selection in instance-based learning approach 251

C.5 Case-studies B and C: Detailed predictive results using an instance-based learning approach based on Kriging 251

References **253**

List of Figures

1.1 Graphical representation of the annual evolution in the number of unique organic and inorganic substances. 2

1.2 Schematic overview of the study objectives. 6

1.3 Schematic general overview of the methodology followed for this study. 10

2.1 Outline of the steps involved in predicting molecular properties from molecular structure in a QSPR/QSAR problem 23

2.2 An example from Mike Hann (1994), inspired by Magritte's painting *"Ceci n'est pas une pipe"*, using an image of a salmeterol molecule. 25

2.3 Different graph tables and matrices for the molecule 1H-Indole-2,3-dione. 32

2.4 Schematic example of the calculation of the molecular similarity/dissimilarity using Fingerprints 43

3.1 Flow chart showing the general steps used to predict properties in the context of QSPR/QSAR problems using model-based approaches. 67

3.2 General workflow of the hybrid approach using Random Forests for Feature Selection for predicting properties of compounds based on molecular descriptors 72

3.3 Proportion of variance in the descriptor set that is explained by each principal component. 78

3.4 Comparison of the RMSE for each predictive SVM model using an increasing number of principal components in descending order of proportion of variance explained (case-study A1). 79
3.5 Boxplots depicting the distribution of the importance score of each variable (case-study A1). 80
3.6 Comparison of the RMSE for each predictive SVM model using an increasing number of variables in descending order of importance (case-study A1). 81
3.7 Comparison of the RMSE for each predictive SVM model using an increasing number of variables in descending order of importance (case-study A2). 85
3.8 Proportion of variance in the descriptor set that is explained by each principal component in case-study D. 89
3.9 Comparison of the RMSE for each predictive SVM model using an increasing number of principal components in descending order of proportion of variance explained in case-study D. 90
3.10 Comparison of the root mean square error (RMSE) for each predictive SVM model using an increasing number of variables in descending order of importance for case-study D. 91
3.11 General workflow of the hybrid approach using Random Forests for Feature Selection for predicting properties of compounds based on molecular descriptors adapted to BBB permeation prediction. . 94
3.12 Boxplots depicting the distribution of the importance score of each variable in case-study G. 95
3.13 Density plots depicting the distribution of the ϕ and overall accuracy with different number of variables for case-study G. 97
3.14 Composite screenshot example of some data retrieval features in the ThermInfo Web information system. 98
3.15 Distribution of the 89 most important variables by classes of descriptors in case-study A1. 102
3.16 Plots for prediction errors analysis of the best model for predicting the properties in case-study A1. 104

3.17 List of the 20 most important variables by classes of descriptors for each property in case-study A2. 106
3.18 Plots for prediction errors analysis of the best models for predicting all properties of the test set in case-study A2. 107
3.19 Density plot showing the distribution and variation of the median EC_{10} in the train and test sets randomly sampled. 109
3.20 Final scoring for NIEHS-NCATS-UNC DREAM toxicogenetics challenge. 110

4.1 Example of four compounds with the same fingerprint 116
4.2 Example of two molecules with the same topology but different topography. 117
4.3 Two small molecules that will used to exemplify the development of NAMS. 128
4.4 Atomic alignment between molecules A and B and atomic and molecular similarity scores. 133
4.5 Basic steps for detecting and classifying a chirality center. 137
4.6 Basic steps for detecting and classifying E-Z isomerism 138
4.7 Distribution and variation of the pairwise similarity between hydrocarbons using daylight fingerprints and NAMS. 142
4.8 Example of the pairwise similarity using fingerprint and NAMS, obtained for 8 compounds extant in the dataset A1. 143
4.9 2D Kernel density estimator perspective and contour plots showing the distribution of the pairwise similarity between the 31 steroids. 144
4.10 The basic skeleton of a steroid and three estrogenic steroids . . . 145
4.11 Fraction of active compounds in the dataset that are similar to seeds with a certain level of activity used for similarity search with different threshold cut-offs. 146
4.12 Fraction of actives compounds within those compounds similar to compounds that are active or inactive using Fingerprints and NAMS with different threshold cut-offs. 147
4.13 Screenshot of the NAMS Web-tool output when comparing two simple molecules. 149

LIST OF FIGURES

5.1 Plots representing the distribution of pairwise distance between pairs of compounds, calculated using *a)* molecular descriptors, *c)* fingerprints and *e)* NAMS and the corresponding absolute difference in the aqueous solubility value. 174

5.2 Plots representing the distribution of pairwise distance between pairs of compounds, calculated using *a)* molecular descriptors, *c)* fingerprints and *e)* NAMS and the corresponding absolute difference in the pIC_{50} value. 178

5.3 Relationship between the selected neighbourhood size and the calculation time. 185

5.4 Density plots of the differences between the observed values and the predicted values for the train and test sets compounds. 187

5.5 Example of the neighbourhood of the test compound *1-122870*. . . 189

5.6 Plots showing the relationship between the maximum similarity between test compounds and compounds of the training set that were selected for predicting the property and predictive performance using Q^2 or RMSE for datasets B and C. 190

5.7 Distance between different examples of training and test compounds versus the pIC_{50} difference for each situation as well as the distribution of the properties in the compounds selected for training, the observed and predicted pIC_{50} of the test compound. 193

5.8 Relationship between true prediction error and estimated predicted error by kriging to assess the quality of prediction. 194

5.9 Effect of the training set size on test set predictive results and maximum similarity between training neighbourhood and test compounds. 196

5.10 Example of a situation (data set B – aqueous solubility) in which the most similar compounds to the test compound (ID: 1259) have considerable differences in the property value. 198

5.11 Example of a situation (data set C – DHFR inhibitors activity) in which the most similar compounds to the test compound (ID: 1-127977) are all highly active, yet the test compound is inactive. 199

6.1 Framework for finding and developing promising compounds with desired target properties. 212

A.1 Density plot showing the distribution and variation of the standard enthalpy of formation in gas phase in the training and testing sets. 223
A.2 Density plots showing the distribution and variation of properties in the ThermInfo dataset. 225
A.3 Density plot showing the distribution and variation of aqueous solubility in the training and testing sets. 227
A.4 Density plot showing the distribution and variation of DHFR inhibition activity in the training and testing sets. 229
A.5 General overview of the data available in NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge 231
A.6 Data Analysis and pre-processing scheme in the NIEHS NCATS UNC DREAM Toxicogenetics Challenge 232
A.7 Density plots showing the distribution and variation of properties in training and testing sets of the NIEHS NCATS UNC DREAM Toxicogenetics Challenge. 233
A.8 Four-fused-ring steroid skeleton. 234
A.9 Density plot showing the distribution and variation of CBG binding affinity data set of steroids. 235
A.10 Diagram representing the implementation of the ELBA parameters' generator. 243

List of Tables

2.1 Different line notations for the molecule 1H-Indole-2,3-dione. . . . 27
2.2 Summary of the most widely used file formats for exchange chemical structure information. . . . 30
2.3 Overview of the most widely used molecule editors/viewers and their details. . . . 31
2.4 Summary the most used software for molecular descriptors calculation. . . . 35
2.5 Fingerprint (1024 bits) and bit set list for the example molecule 1H-Indole-2,3-dione. . . . 37
2.6 The most widely used metrics to compare chemical molecules based on fingerprints. . . . 42

3.1 Summary of the results (10-fold cross validation) obtained for all the models (case-study A1). . . . 77
3.2 Summary of the results (10-fold cross validation) obtained for all the models and properties of case-study A2. . . . 84
3.3 Summary of the results (10-fold cross validation) obtained using the best model for each property in an independent test set of case-study A2. . . . 86
3.4 Summary of the results (10-fold cross validation) obtained for all the models for case-study D. . . . 88
3.5 Summary of the results obtained for BBB penetration modelling. 96

4.1 Sample similarity values (V_{nd}) between *aba-bonds* extant in Molecule *A* . . . 129

4.2 *Aba-Bonds* topological distances (d) starting from each atom α_i of the molecule A. 130
4.3 Example of demoninators for bond comparison. 131
4.4 Example of *aba-bond* scores for two atoms of molecule A. 131
4.5 Atom similarity scores between the example molecules A and B. . 132
4.6 The weights (W_m) of all characteristics under consideration for the similarity calculation using NAMS. 139

5.1 Summary of the best results obtained for training set(dataset A1) using different neighbourhood selection methods. 172
5.2 Comparison of the predictive power of the model developed in this study with other model-based approaches for case-study A1. . . . 172
5.3 Summary of the best results obtained for training and testing sets (dataset B) using each dissimilarity matrix. 175
5.4 Comparison of the predictive power of the model developed in this study with other published models with the best results for the same dataset. 177
5.5 Summary of the best results obtained for training and testing sets (dataset C). 180
5.6 Comparison of the predictive power of the model developed in this study with other published models with the best results for the same dataset. 183
5.7 Summary of the highest prediction errors using the best model to predict enthalpy of formation of gas phase (case-study A1). 186

A.1 An overview of the case-studies. 219
A.2 Distribution of the compounds in the training and independent validation sets into the different types of hydrocarbons. 222
A.3 Distribution of the compounds in the training and testing sets for six thermochemical properties in the ThermInfo dataset. 224

Chapter 1

Introduction

1.1 Context: Data Explosion in Chemistry

Chemoinformatics - A new name for an old problem?
~ M. Hann and R. Green (1999)

Large scale research projects are becoming part of chemistry research in more and more laboratories, producing an ever-increasing amount of data and information. The chemo-information keeps growing exponentially due to constantly refined and optimized experimental technologies (Bachrach, 2009; Chen, 2006). According to Chemical Abstracts Service (CAS) [1] there are currently more than 77 million known substances of which 10 million were added in less than one year (Figure 1.1). This database is updated daily and approximately 12,000 new substances are added each day. In comparison, it took 33 years for CAS to register its first 10 million substances in 1990, which is an indicator of the accelerating pace of chemical knowledge (Figure 1.1) [2].

Thus, it was realized that the amount and complexity of information accumulated by chemists can only be managed by exploring it using computer technologies (Bajorath, 2004; Gasteiger, 2003). This problem led to a new field of expertise – the intersection of chemistry and computer science, with emphasis on the acquisition, manipulation, organization, analysis and dissemination of

[1] Chemical Abstracts Service: http://www.cas.org/, accessed in December, 2013

[2] Data from "CAS Statistical Summary 1907-1997," Chemical Abstracts Service, Columbus, Ohio: http://www.shinwon.co.kr/cas/ASSETS/casstats.pdf, accessed in December, 2013

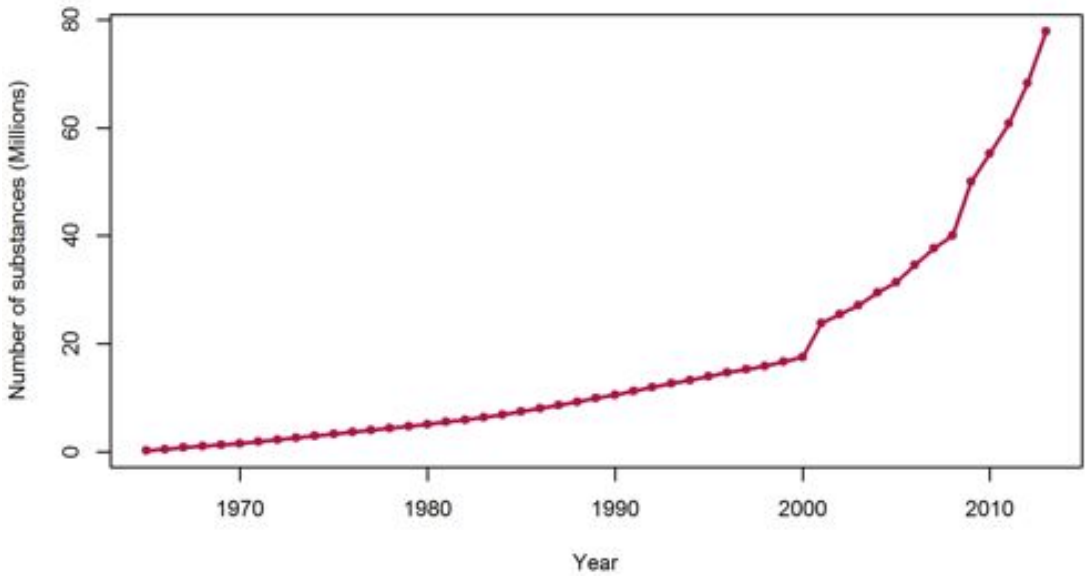


Figure 1.1: Graphical representation of the annual evolution in the number of unique organic and inorganic substances recorded in the Chemical Abstract Service Registry System [1, 2] between 1965 and 2013.

chemical data and information (Bajorath, 2004; Chen, 2006). This field of expertise clearly spans a very large (and still to be defined) range of problems and approaches and it does no longer imply, as it did in the beginning, that it is necessarily related to drug discovery. This new interdisciplinary area was named *"Chemoinformatics"* by Brown & James (1998). In this article, chemoinformatics is defined, yet very focused on drug discovery process, as follows:

"The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization".

As for the definition of chemoinformatics, the name of the area is not universally agreed upon (Bajorath & Warr, 2011). In the literature, several terms are used synonymously to chemoinformatics: cheminformatics, chemi-informatics, chemical informatics, chemometrics, computational chemistry, etc (Bajorath, 2004).

A Google search, in December 2013, retrieved ∼467,000 hits for the term *"cheminformatics"* whereas *"chemoinformatics"* had ∼261,000 hits. The results indicate that cheminformatics is more commonly applied and therefore it will be used throughout the document.

Chemoinformatics has different practical applications in different areas such as pesticide, drug and material design, environmental protection, food safety, among others.

1.2 Motivation

Chemistry is (almost) everywhere and in everything.
∼ A. Shani (2004)

"When you hold this document you are holding molecules. When you drink coffee you are ingesting molecules, as you sit in a room you are bombarded by a continuous storm of molecules. When you appreciate the colour of an orchid and the textures of a landscape you are admiring molecules. When you savour food and drink you are enjoying molecules. When you sense decay you are smelling molecules. You are clothed in molecules, you eat molecules, and you excrete molecules. In fact, you are made of molecules" (Atkins, 2003). In other words, molecules are (almost) everywhere and in everything, and as mentioned above the number of molecules discovered each day continues to grow at an exponential rate due to constantly refined and optimized experimental technologies (Bachrach, 2009). However, the experimental determination of the chemical, physical and biological properties (from this point on simply referred as properties) of compounds is often expensive, time-consuming and in many cases impossible. According to George Hammond in the 1968 Norris Award Lecture, *"the most fundamental and lasting objective of synthesis is not production of new compounds, but production of properties"*, thus it is evident that there is a great need to organize and make high quality experimental data available to the scientific community and foster the application of property prediction methods with a good predictive performance when experimental values are not available which is essential to many industries and technologies. One of the most promising areas in cheminformatics is the development of methods aimed at predicting these properties from the

structure of the molecules. Unlike quantum chemistry or molecular simulation, which are designed to model physical reality, cheminformatics is intended simply to produce useful models that can predict properties of compounds given their structure. These methods are usually known as Quantitative Structure-Property/Activity Relationship (QSPR/QSAR). During the last twenty years QSPR/QSAR have been applied to a wide range of problems gaining an extensive recognition in physical, organic, analytical, pharmaceutical and medicinal chemistry, biochemistry, chemical engineering and technology, toxicology, and environmental sciences (Micheli, 2003). Examples of the wide range of predicted properties include melting and boiling temperature, molar heat capacity, vapor pressure, solubility, viscosity and partition coefficients, standard enthalpy of formation, refractive index, density, solvation free energy, receptor binding affinities, pharmacological activities, and enzyme inhibition constants (Micheli, 2003).

It is important to understand that QSPR/QSAR models will not replace experimental measurements, however they offer multiple advantages with an enormous scientific, humanitarian and economic impact: (1) innovation, to analyse manually such a huge amount of chemical data is obviously impossible, and thus computer, *in silico* methods, will represent a privileged way to explore, discover and design promising compounds with desired properties; (2) prioritizing needs by selecting the most promising untested and sometimes yet unavailable compounds; (3) reduction of time needed for experiments as they serve as a filter to reduce the number of compounds that need to be tested; (4) even in an hypothetical situation of trying to experimentally study all properties of all compounds, the amount of existing laboratories and human resources is not sufficient to deal with this quantity of chemical data; (5) reduce the costs by reducing the number of measurements as the cost of performing experimental measurements is in most cases very high; (6) reduction of the number of animals needed for *in vivo* experiments which is ethically and economically very important; (7) the development of new chemicals is often centred on the target properties for the candidate new product, however critical issues, such as toxicity, industrial safety and environmental health should also be evaluated from the beginning - reducing the economic resources needed to the development of chemicals without the knowledge of their toxicological and environmental properties. As an example, various

studies estimate that traditional development of a new prescription drug takes between 10 and 20 years and costs an average of $500-$800 million (Dickson & Gagnon, 2009). Consequently, the burden to reduce costs and accelerate drug discovery is high, especially considering the human benefits of such achievement and the constant inevitability to cope with new diseases. Besides the efficiency of the process, it is also important to consider the need to improve the percentage of compounds with high therapeutic value and to reduce the side effects of drugs. The advantages of the computational methods extend to the primary and earlier stage of the complex drug discovery process: drug lead identification and elimination of compounds that are toxic or have poor pharmacokinetic properties.

1.3 Problem Statement and Aims of the Study

While much bioscience is published with the knowledge that machines will be expected to understand at least part of it, almost all chemistry is published purely for humans to read.
~ P. Murray-Rust et al. (2004)

Various analytical tools from statistics and machine learning are used in QSPR/QSAR analysis including predictive modelling (classification and regression), visualization, exploratory data analysis and cluster analysis. These studies rely on the principle that states that similar compounds tend to have similar properties (Johnson & Maggiora, 1990). The fact that the domain of QSPR/QSAR problem is naturally composed by unstructured data, as molecules can have arbitrary dimension, structure and composition, and the fact that there is not a univocal and unequivocal way of coding and comparing these molecules make it challenging to apply machine learning techniques. Several approaches exist and several have provided good results for specific domains, however, to the best of our knowledge, one cannot expect a QSPR/QSAR approach to work well to predict any property, the set of descriptors that allows predictions with good predictive power depend highly on the property of interest and most methodologies work like "black boxes" without a detailed understanding of each prediction and expected prediction error.

Having in mind all the strengths and limitations of the existing databases and prediction methods, the thesis underlying the present work is that *it is possible*

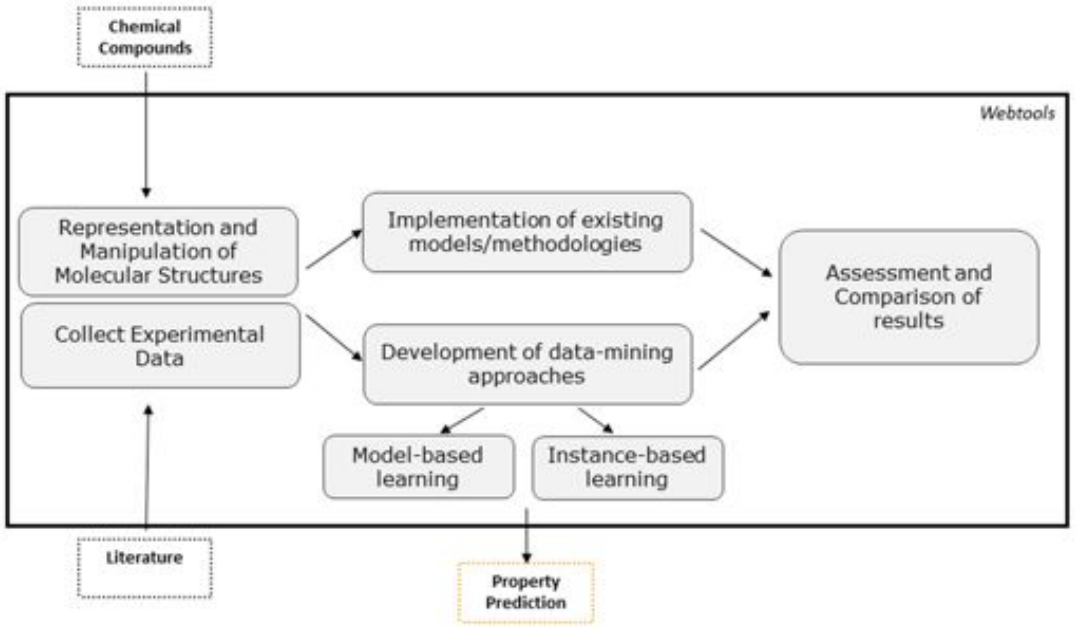


Figure 1.2: Schematic overview of the study objectives. **a)** Representation and manipulation of molecular structures and experimental data. **b)** Development of data-mining models. **c)** Implementation of existing models. **d)** Assessment and comparison of results. **e)** Development of Web-based systems to disseminate the results.

to improve the current models for the prediction of physical, chemical and biological properties based solely on the chemical structure using advanced automated analysis solutions based on Machine Learning. The aims of the study cover the development and implementation of cutting-edge machine learning and statistical modelling algorithms for handling large-scale chemical data in order to improve the prediction of properties not only in terms of predictive power but also improving the robustness and comprehensibility of such methodologies. The more specific aims of this study are represented in Figure 1.2 and include:

- Compile and make available good experimental data of chemical, physical or biological properties of molecules, since without good data it is not possible to develop good predictive models (i. e. *"garbage in, garbage out"*);
- Study the theoretical basis of QSPR/QSAR modelling and the machine

learning methods used in this field;

- Understand the specificities of the representation of chemical structures in computer readable formats which is required for data analysis. Physical, chemical as well as biological properties are in large part determined by the molecular structure. There are several ways to represent a molecular structure and different representations contain different chemical information. One of the major tasks in automated extraction of meaning, patterns, and regularities using machine learning methods is to represent chemical structures, to transfer the various types of chemical information taking into account their complex and heterogeneous nature into a machine-readable representation that can be processed by a machine learning model. Hence, it is important to select machine-readable representations and machine learning models that can handle and extract the right chemical data according to the chemical property that needs to be predicted;

- Implement and assess existing prediction models with experimental data extracted from several sources verifying the quality of results produced and develop and validate new model-based machine learning approaches to improve the results;

- Implement and assess the most widely used methodologies to calculate similarity between molecules;

- Develop a new algorithm to adequately quantify the structural similarity between molecules with an high discriminative power of similar molecules.

- Develop and assess an instance-based method that, in light of the structural similarity principle (Johnson & Maggiora, 1990), takes into account the high dimensionality of the chemical space, predicting chemical, physical or biological properties based on the most structurally similar compounds in the molecular space, consequently avoiding the selection of descriptors, increasing the robustness and comprehensibility of the method;

- Develop and implement methods that automatically search the neighbourhood of a compound and determine the optimal number of neighbours which can be used to predict its property with a minimized prediction error;
- Design and develop experiments to be simultaneously proof-of-the-concept and applicable to existing experimental data. In fact, besides investigating the methodology and developing new models, one of the main aims is to investigate the impact of the approach on methods developed for real-world data and problems;
- Design and implement tools as well as make publicly available to the community open source code that not only allows access to data in a comprehensive way but also permits using methodologies developed in the context of this work.

The research questions which guide the development of this work are:

- Can chemical, physical and biological properties prediction be improved in terms of prediction error, robustness and comprehensibility using a new descriptor selection technique coupled with a model-based machine learning algorithm?
- Can quantification of structural similarity be improved using an algorithm that is based on atom matching?
- Can chemical, physical and biological properties be predicted by an instance-based machine learning approach using as input a metric space constructed based on structural similarity?
- Does an instance-based machine learning approach using structural similarity to construct a metric in order to predict chemical, physical and biological properties has advantages in terms of predictive performance, robustness and comprehensibility in relation to a model-based machine learning approach?
- Is it possible to increase the predictive results and comprehensibility of the method using smaller local neighbourhoods?

1.4 Methodology Overview

Why,' said the Dodo, 'the best way to explain it is to do it.'
~ Lewis Carroll in *Alice in Wonderland*

This work aims to increase the capability to predict physical, chemical and biological properties, using data mining methods applied to complex non-homogeneous data (chemical structures), for large information repositories. For that purpose we explore two avenues where machine learning can help (Figure 1.3), (1) building predictive models using model-based approaches that will establish a quantitative relationship between structure and property and (2) using instance-based approaches which use the most similar compounds to interpolate properties, effectively using a richer and comprehensive hypothesis space to form an implicit global approximation to the target function.

Current methodologies in QSPR/QSAR attempt to relate a set of selected structure-derived features of a compound to its property using model-based learning. For that purpose there are three main steps that were considered important to follow: (1) select machine-readable structure representations and machine learning models that can handle chemical data; (2) implement and assess some existing prediction models with experimental data extracted from several sources verifying the quality of the results produced; (3) implement Web tools that not only allow access to the experimental data in a comprehensive way but also permit estimating properties/activities using different prediction methods without requiring previous knowledge about those methods. This research focused on major issues identified when predicting properties of chemical compounds and on the solutions explored using different molecular representations, feature selection techniques and data mining approaches. The advantages, disadvantages and specificities of molecular representations, feature selection techniques and data-mining techniques for prediction of chemical, physical and biological properties using different *real-world* datasets (Appendix A) were explored. It was found that one of the main requirements in the development of these QSPR/QSAR predictive models is the identification of the subset of variables that represent the structure of a molecule and which are predictors for a given property. The

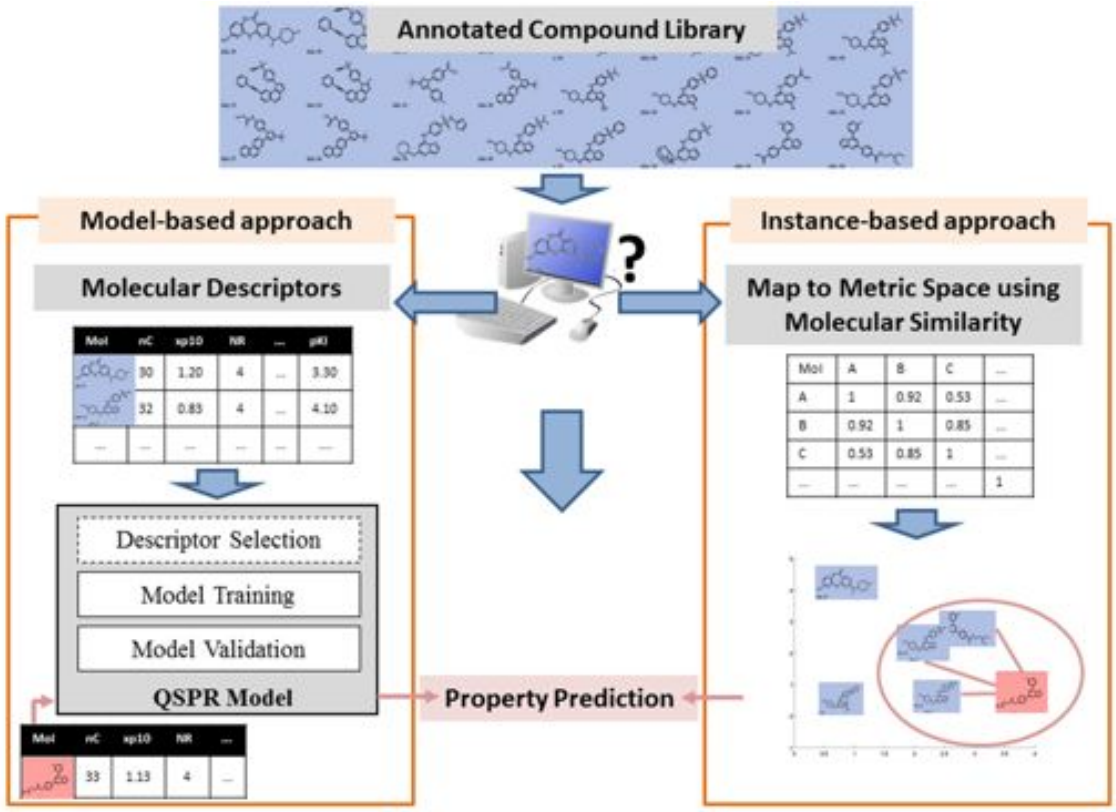


Figure 1.3: Schematic general overview of the methodology followed for this study to predict chemical, physical and biological properties: given a set of compounds for which their properties are known two types of modelling approaches were followed: **a)** Model-based approaches, which use molecular descriptors to represent molecules, followed by selection of subsets of descriptors and model development and validation which stablish a quantitative relationship between structure and property. To predict properties of new compounds for which properties are unknown, their structure should be represented using the same subset of descriptors defined in the training phase and then apply the fitted model. **b)** Instance-based approaches, which represent the molecules in the chemical space using the similarity or dissimilarity between them and use an interpolation technique that is able to estimate properties based on the instance space. To predict properties of new compounds for which properties are unknown, they should be represented in the chemical space, a neighbourhood of training compounds should be selected and the interpolation technique should be applied using the compounds in this neighbourhood.

problem lies in selecting the minimum subset of descriptors that can predict a certain property with a good performance, computationally more efficient and in a more robust way, since the presence of irrelevant or redundant features can cause poor generalization capacity. An alternative selection method, based on Random Forests to determine the variable importance was developed in the context of QSPR/QSAR problems. The subsequent predictive models are trained with support vector machines introducing the variables sequentially from a ranked list based on the variable importance (Chapter 3). The models were then validated both internally and externally.

Given that similar molecules tend to have similar properties, we alternatively developed an instance-based machine learning methodology for predicting properties of compounds using the similarity-based molecular space (Chapter 5). This new approach takes into account the high dimensionality of the molecular space, predicting chemical, physical, or biological properties based on the most similar compounds (neighbourhood) with measured properties. Different approaches to select the best neighbourhood to predict properties were also designed and explored. This methodology uses ordinary kriging coupled with molecular similarity approaches which creates an interpolation map over the molecular space that is capable of predicting properties/activities for diverse chemical data sets. However, as already mentioned this type of methodology requires the quantification of structural similarity between molecules, which is often subjective, ambiguous and relies upon comparative judgements, and consequently, there is currently no absolute standard of molecular similarity. In this context, we developed a new similarity method, the Non-contiguous Atom Matching Structural Similarity function (NAMS), based on an iterative directed graph similarity procedure and optimal atom alignment using pairwise matching algorithms (Chapter 4). In general, to solve the global problem of quantifying the structural similarity between molecules, we decided to break it down into solvable different parts by reducing the molecule to atoms and compare atoms of different molecules in order to find the best alignment between them. These atoms are considered not only in terms of their intrinsic chemical characteristics but also according to their relation to the other atoms in the molecule. The similarities detected by an atom correspondence approach like the present one are consistent with the chemistry and

structure of molecules because it depends on the direct neighbourhood of each atom as well as the overall topology of the molecule, becoming more intuitively understood because similar atoms in molecules are explicitly shown.

Data and methodologies developed in the context of this work were made available to the scientific community in tools that were implemented with the objective of disseminating the results and at the same time allowing access to data and methodologies collected and developed in the context of this work.

1.5 Contributions and Publications

Conducting data analysis is like drinking a fine wine. It is important to swirl and sniff the wine, to unpack the complex bouquet and to appreciate the experience. Gulping the wine does not work.

∼ Daniel B. Wright (2003)

As result of the work presented in this dissertation, the following contributions and resulting journal and conference publications can be highlighted [1]:

- Explored advantages, disadvantages and specificities of molecular representations and data-mining techniques for prediction of chemical, physical and biological properties using different *real-world* datasets (Appendix A) and developed an innovative approach to improve the prediction power and comprehensibility of QSPR/QSAR problems using Random Forests for feature selection. This methodology seemingly improves the prediction performance of the models using a limited set of molecular descriptors, providing faster and more cost-effective calculation of descriptors by reducing their numbers, and providing a better understanding of the underlying relationship between the molecular structure represented by descriptors and the property of interest (Chapter 3):

    1. Ana L. Teixeira, João P Leal, Andre O Falcao 2013: **Random forests for feature selection in QSPR Models - an application for**

[1] Publication list with web-links to the contents is available at: http://xldb.di.fc.ul.pt/wiki/Ana_Teixeira

predicting standard enthalpy of formation of hydrocarbons. Journal of Cheminformatics, 5:9.

2. Participated in the 8th Dialogue on Reverse Engineering Assessment and Methods (DREAM 8) Challenges (June – September, 2013) NIEHS-NCATS-UNC Toxicogenetics Challenge (229 Participants organized in 24 teams) to model population-level cytotoxicity parameters to unknown pharmaceutical chemical compounds based on chemical structure attributes. The model that was submitted ranked first position assessed using as criteria the root mean squared error and second position assessed using as criteria the root mean squared error and Pearson and Spearman correlation coefficients).

3. Participated as part of the consortium elaborating a manuscript describing an analysis of the different methods and results that arise from this challenge to be published in Nature Biotechnology in the near future.

- Developed an Information System – *ThermInfo* - to Collect, Retrieve, and Predict Thermochemical Data, this system integrates a database for structural and thermochemical properties of organic compounds and a user-friendly Web-interface that is publicly available to a broad community which not only allows access to the data in a comprehensive way but also permits estimating properties using different prediction methods (namely, the Extended Laidler Bond Additivity (ELBA) method (Santos *et al.*, 2009) and methods developed in the context of this work) without requiring previous knowledge about those methods (Chapter 3):

  1. Ana L. Teixeira, Rui C. Santos, João P. Leal, José A. Martinho Simões, Andre O. Falcao, **ThermInfo: Collecting, Retrieving, and Estimating Reliable Thermochemical Data**. Technical Report. Department of Informatics, Faculty of Sciences, University of Lisbon, January 2013.

  2. ThermInfo Web-tool: `http://therminfo.lasige.di.fc.ul.pt`

- Developed an algorithm for automatic identification and classification of molecular stereochemistry, a key step for the assessment of structural similarity (Chapter 4):

    1. Ana L. Teixeira, João P. Leal, Andre O Falcao. **Automated Identification and Classification of Molecular Stereochemistry: Chirality and Double Bond Stereoisomerism**. Technical Report. University of Lisbon, Faculty of Sciences, LaSIGE, 2013.

- Developed an innovative structural similarity method (Noncontiguous Atom Matching Structural Similarity function (NAMS)) based on atom alignment between both molecules and assessed it against widely used similarity functions. Despite having a higher computational cost, the method performs well being able to distinguish either different or very similar compounds that were indistinguishable using a fingerprint-based approach. NAMS also verifies the similarity principle by showing that pairs of compounds with a high degree of similarity tend to have smaller differences in the absolute value of their property. The method is also able to recover a significantly higher average fraction of active compounds when the seed is active for different cut-off threshold values of similarity (Chapter 4):

    1. Ana L. Teixeira, Andre O. Falcao, 2013: **Noncontiguous atom matching structural similarity function**. Journal of Chemical Information and Modeling, 53 (10), pp 2511–2524.

- Designed and implemented a Web-tool that makes NAMS available for the community in a *user-friendly* way as well as its source code as a Python package:

    1. NAMS Web-tool: `http://nams.lasige.di.fc.ul.pt`
    2. NAMS-0.9.2: `https://pypi.python.org/pypi/NAMS/0.9.2`

- Developed a new method to predict physical, chemical or biological properties of molecules using molecular structural similarity and Ordinary Kriging.

The overall predictive results using kriging comply with the results obtained in the literature using typical QSPR/QSAR approaches. However, the procedure did not involve any type of descriptor selection or even minimal information about each problem, suggesting that this approach is directly applicable to a large spectrum of problems in QSPR/QSAR. Furthermore, the predictive results improve significantly with the similarity threshold between the training and testing compounds, allowing the definition of a confidence threshold of similarity and error estimation for each case inferred. The use of kriging for interpolation over the molecular metric space is independent of the training data set size, and no re-parametrizations are necessary when compounds are added or removed from the set, and increasing the size of the database will consequentially improve the quality of the estimations. This methodology can also be used for checking the consistency of measured data and for guiding an extension of the training set by determining the regions of the molecular space for which new experimental measurements could be used to maximize the predictive performance of the method (Chapter 5):

1. Ana L. Teixeira, Andre O Falcao, 2014: **Structural similarity based kriging for quantitative structure activity and property relationship modeling**. Journal of Chemical Information and Modeling 54(7), 1833-1849.
2. Andre O Falcao and Ana L. Teixeira. **Method and apparatus for quantitative prediction of physical, biological or pharmacological properties of molecules using molecular structural similarity**. Provisional Patent Application number 107361 (filing date December, 2013)

- Collaborated in the development of new approaches based on Bayesian statistics coupled with machine learning and feature selection methods to produce robust models in real-world drug research scenarios (Chapter 3):

1. Ines Filipa Martins, Ana L. Teixeira, Luis Pinheiro, Andre O. Falcao 2012: **A Bayesian Approach to *in Silico* Blood-Brain Barrier**

Penetration Modeling. Journal of Chemical Information and Modeling 6(52), 1686-1697.

2. Ana L. Teixeira, Andre O Falcao. **Improving Blood Brain Barrier Penetration *in-silico* Models With a Hybrid Approach for Descriptor Selection**. Workshop Beating the Blood-Brain and Other Blood Barriers, Lisbon, February 2013.

- Collaborated in the development of an Information System for Blood-Brain Barrier penetration data (*B3Info*) and prediction (*B3PP*) (Chapter 3):

1. Andre O Falcao, Luis Pinheiro, Ana L. Teixeira. **B3Info – An information system for molecular Blood-Brain Barrier penetration data**. Workshop Beating the Blood-Brain and Other Blood Barriers, Lisbon, February 2013.
2. B3Info Web-tool: http://b3info.lasige.di.fc.ul.pt
3. B3PP Web-tool: http://b3pp.lasige.di.fc.ul.pt

1.6 Overview of the Document

We are drowning in information and starving for knowledge.
~ Rutherford D. Roger (1985)

The rest of the document is organized as follows:

- Chapter 2 gives a background on several concepts to explain this work and gives an overview of the existing property prediction methods based on structure of molecules. The chapter starts with an introduction to QSPR/QSAR studies, its applications and limitations. The components involved in the construction of QSPR/QSAR are also presented, namely computer-readable representation of molecules, selection of relevant descriptors, machine learning approaches to establish mathematical relationships between structure and property and approaches to assess and validate these models.

- Chapter 3 introduces a methodology developed for model-based learning and the most relevant results are presented and discussed using different case-studies and different sets of descriptors.
- Chapter 4 describes a fundamental problem in cheminformatics, the definition of a new structural similarity algorithm - the Non-contiguous Atom Matching Structural Similarity (NAMS) function. NAMS is the first step in the development of algorithms for property prediction based on instances. The methodology is described in detail as well as its application to different case-studies as compared to widely used structural similarity methodologies.
- Chapter 5 presents an instance-based learning modelling methodology which is based on the kriging algorithm that requires the use of the chemical space based on the distance between molecules and the results obtained using different case-studies are presented and discussed.
- Chapter 6 revisits the thesis and objectives raised earlier in this chapter, expresses conclusions in the light of the contributions obtained with this work and discusses future directions of the present work.
- Appendix A is dedicated to the description of the case-studies used in this work, collection and cleaning of datasets of experimental properties as well as the selection and preprocessing of different molecular representations which are assumed to influence the property of the molecule.
- Appendix B makes available the datasets of experimental properties as well as descriptor sets used in this work.
- Appendix C provides detailed results obtained for different case-studies and methodologies developed and presented in this work.

Chapter 2

Background

Chemists have always employed models aimed at representing complex chemical entities in a simple and comprehensible way: names, molecular weight, graphical draws, and so on. But the rising of computer science has allowed developing a great amount of methods with the aim of transforming molecules into data structures amenable to be processed by computers. Machine learning and knowledge exploration of chemical information are key in research areas such as drug discovery, property/activity prediction, modelling of structures, and many others, where the meaningful linking of experimental knowledge and information about chemical structure is necessary. These applications are often based on quantitative structure activity/property relationship (QSPR/QSAR) models, where relevant information for models is extracted from large data sets of molecular descriptors. Several classifications of these methodologies could be carried out depending on the characteristics of the molecular descriptors (for example, 2D or 3D, local or global, constitutional or geometrical), the selection of the most important descriptors for the property in study and on the multivariate method employed to define the QSPR/QSAR equation (for example, model- or instance-based learning). These approaches are based on the principle that states that *"structurally similar molecules show similar properties"* (Johnson & Maggiora, 1990). This chapter discusses the development of QSPR/QSAR and lists the applications and limitations of these methodologies. The components involved in the construction of QSPR/QSAR are presented, namely computer-readable representation of molecules, selection of relevant descriptors, machine learning

approaches to establish mathematical relationships between structure and property and approaches to assess and validate these models.

2.1 Quantitative Structure-Property/Activity Relationship Methods

"Measure what is measurable, and make measurable what is not so."
∼ Galileo Galilei (1564-1642)

The determination of QSPR/QSAR have been applied in chemistry and related research areas for several decades. The QSPR/QSAR approach can be generally described as an application of machine learning methods and statistics to develop models that could accurately predict chemical, physical or biological properties of compounds based on their structures, assuming that there is a strong correlation between them (Dudek *et al.*, 2006). This QSPR/QSAR modelling approach is underpinned by the so-called *principle of similarity*, that chemical compounds with similar chemical structures will have similar activities (Johnson & Maggiora, 1990). This assumption is a prerequisite both for the meaningful description of trends within the training set, and for the interpolation of those trends to encompass other compounds. An application of this fundamental assumption requires that the notion of similarity of chemical structures be made precise. Formulating this measure of structural similarity is itself intimately bound with how the chemical structure is numerically represented which is still a challenge in cheminformatics. Furthermore, reliable experimental data are required to build reliable predictive models, with a clear and unambiguous endpoint (Tropsha & Golbraikh, 2007).

The first developments of the methodology date back to 1868 when Brown & Fraser (1868) formulated the following hypothesis: *"a relation exists between the physiological action of substance and its chemical composition and constitution, understanding by the latter term the mutual relations of the atoms in the structure"*. Followed by pioneering studies of Mills (1884), Meyer (1899), and Overton (1901) which separately found linear relationships between structure and properties/biological effects. QSPR/QSAR progressed during the 1930s to 1960s after

studies by Hammett (1935), Taft Jr (1952) and Hansch & Fujita (1964). Since then, the methodology has been increasingly used in drug discovery (e.g. Ekins *et al.* (2007); Kubinyi (1997a); Lill (2007); Lipinski *et al.* (2012); Martin (1998); Perkins *et al.* (2003)) and chemical technology (e.g. Du Xihua & Keying (2009); Katritzky & Fara (2005); Katritzky *et al.* (2000)). These studies received a big boost with the application of computers and increase of the computing power. This has not only led to the proposition of newer and more complex molecular representations, but also in the application of prediction techniques that were either not feasible or were previously too time consuming. Today the methodology has become interdisciplinary, with an extensive number of available tools for generating and harvesting information about chemical structure and linking this structural information with experimental measurements of properties or activities using machine learning algorithms in order to extract new knowledge.

The three major difficulties in the development of QSPR/QSAR models are quantifying the inherently abstract molecular structure, determining which structural features most influence the given property (representation problem) and then establishing the functional relationship that best describes the relationship between these structure descriptors and the property data (mapping problem) (Kubinyi, 1997b). The first difficulty is often overcome by calculating molecular descriptors, which are developed to quantify various aspects of molecular structure. In fact, this solution is the cause of the second difficulty since, as described in the following sections, hundred of molecular descriptors exist describing a wide range of constitutional, topological, geometrical, electronic and quantum mechanic features. These descriptors may in turn be highly redundant, since many descriptors are related to each other or to the same underlying property. Furthermore, some descriptors may be completely irrelevant from the desired property's point of view, and others may have been calculated with methods producing noisy values. The problem lies in the identification of the appropriate set of descriptors that allow the desired property of the compound to be adequately predicted. To accomplish this and to find the optimum relationship between these structure descriptors and the property data, several statistical and machine learning methods are used for dimensionality reduction or feature selection and regression or classification. Models can be grouped into two main categories depending on the

nature of the property to be predicted. Models predicting quantitative properties, such as the degree of binding to a target, are known as regression models. On the other hand, classification models predict qualitative properties. However, there are also difficulties in the modelling phase, namely the properties used to build models often originate from complicated and uncertain measurements, resulting in noisy y-values. The values may also have been collected from different public sources with varying reliability or obtained in different experimental conditions. Another common problem is the unbalanced nature of the available data, that is, the majority of the compounds in a database are inactive, whereas only a few compounds are active or *vice-versa*. Even if all descriptors and output are measured and calculated as accurately as possible, it is still problematic to make good models. Particularly difficult are compounds in a chiral pair. The two isomers have identical attributes, but very often completely different activities — one isomer might be toxic and the other one might not. This phenomenon of molecules having essentially different properties though very similar structure is generally known as *activity cliff* (Stumpfe & Bajorath, 2012).

In order to build the model, the pool of molecules with known activity is usually split into a training set and a test set. The training set is used to learn the model. The learning problem consists in constructing a model that is able to predict properties of molecules in the training set, without over-learning it. Choosing a model among the profusion of existing models is related to the final goal of the study, and while complex models can for instance have a great predictive ability, this often comes in detriment of their interpretability and overfitting. The overfitting phenomenon can for instance be controlled using different validation techniques (described below) that quantify the ability of the model to predict a subset of the training set that was left out during the learning phase. The test set is used to evaluate the generalization of the learned model, corresponding to its ability to make correct prediction on a set of unseen molecules (Tropsha, 2010). This is a key step in QSPR/QSAR modelling, as pointed out by Truchon & Bayly (2007), the major reason why these models fail is attributed to the vast number of equivalent models and deficient external validation. In other words, it is because model overfits the training data without detecting the true structure-activity relationship. Furthermore, QSPR/QSAR models retain a limited scope

of application (Jaworska *et al.*, 2005). The uncertainty and variance are expected for predictions made beyond the scope.

The process of a general QSPR/QSAR problem is summarized in Figure 2.1. The flowchart shows the fact that a QSPR/QSAR model is an alternate path to the prediction of molecular properties since its direct calculation is generally not feasible.

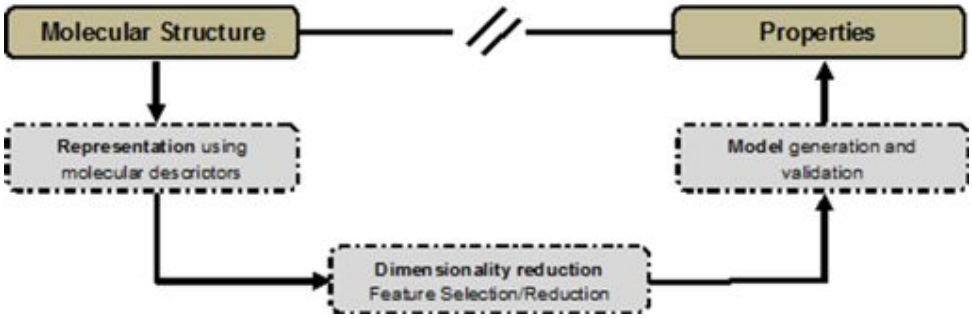


Figure 2.1: Outline of the steps involved in predicting molecular properties from molecular structure in a QSPR/QSAR problem.

Several statistical and data-mining techniques have been employed and software incorporating all the workflow for the determination of QSPR/QSAR as a *black-box* has been created, the vast majority of these being available on a commercial basis only (Baumann *et al.*, 2008; Gasteiger, 2003). ADAPT (Automated Data Analysis and Pattern Recognition Toolkit)[1] is a commercial software system for UNIX operating system distributed by Jurs for the development of QSPR/QSAR. It implements an inductive approach where the QSPRs or QSARs are developed from a set of known values for compounds in a training set. ADAPT has a large selection of molecular structure descriptor generation routines (Stuper & Jurs, 1976). The commercial computer program PASS (Prediction of Activity Spectra for Substances)[2] developed by the Academy of Medical Sciences, Moscow, predicts biological activity for a compound on the basis of its structural formula using Multilevel Neighbourhoods of Atoms (MNA) and Quantitative Neighbourhoods of Atoms (QNA) descriptors (Lagunin *et al.*,

[1]ADAPT: http://research.chem.psu.edu/pcjgroup/adapt.html
[2]PASS: http://www.pharmaexpert.ru/PASSOnline/

2000). The CODESSA[1] commercial software combines a large variety of classical non-empirical molecular descriptors together with more novel quantum chemical and combined descriptors, derived solely from the molecular structure, and invokes both standard and advanced statistical data treatment techniques for the development of QSPR/QSAR correlations in very large descriptor spaces. OpenMolGRID[2] is a free software implementing data mining techniques used for the development of predictive models for estimating various chemical properties and biological activities. OpenMolGRID system provides a flexible infrastructure for automating this kind of scientific workflows. OpenMolGRID system has Grid adapters for several existing software packages that are required for carrying out tasks in the QSPR/QSAR model development workflows (Darvas *et al.*, 2004). Chembench is a free web-based tool for QSPR/QSAR modelling and prediction. The Chembench[3] provides tools for data visualization and embeds a workflow for creating and validating predictive QSPR/QSAR models (Walker *et al.*, 2010). In addition to specific software to QSPR/QSAR correlations, several general statistics or data-mining software can be used for the same purpose, such as SAS[4], SPSS[5], STATISTICA[6], MatLab[7], R[8] and Weka[9].

2.2 Approaches to represent chemical structures

"Formal symbolic representation of qualitative entities is doomed to its rightful place of minor significance in a world where flowers and beautiful women abound."
∼ Albert Einstein, "Hyperbolic Aesthetic"(1937)

In the *Treachery of Images*, René Magritte painted a realistic rendition of an ordinary object, a pipe and written below in script: *"Ceci n'est pas une pipe"* (in English, this is not a pipe) (Figure 2.2). This seems a like contradiction, but it

[1]CODESSA: http://www.semichem.com/codessa/default.php
[2]OpenMolGRID: http://www.openmolgrid.org/
[3]Chembench: http://chembench.mml.unc.edu
[4]SAS: http://www.sas.com/
[5]SPSS: http://www-01.ibm.com/software/analytics/spss/
[6]STATISTICA: http://www.statsoft.com/Products/STATISTICA
[7]MatLab: http://www.mathworks.com/products/matlab/index.html
[8]R: http://www.r-project.org/
[9]Weka: http://www.cs.waikato.ac.nz/ml/weka/

is actually true, the image is not a pipe (*"just try to fill it with tobacco"*), but a depiction or a representation of a pipe (Spitz, 1994). Inspired by Magritte's painting Mike Hann uses an image of a molecule and below in script: "Ceci n'est pas une molécule" (in English, this is not a molecule) (Figure 2.2), it *"serves to remind us that all of the graphics images presented are not molecules, not even pictures of molecules, but pictures of icons which we believe represent some aspects of the molecule's properties"*.

Figure 2.2: An example from Mike Hann (1994), inspired by Magritte's painting *"Ceci n'est pas une pipe"*, using an image of a salmeterol molecule.

One of the major tasks in cheminformatics is to represent chemical structures and to transfer the various types of representation into computer-readable formats. Compared with other scientific disciplines that only use text and numbers for representing data, chemistry has a special challenge: molecules. Molecules consist of atoms held together by covalent chemical bonds. Moreover, molecules can be transformed into other molecules by chemical reactions. Therefore, chemical information not only comprises text and numbers but also has to characterize chemical compounds with their special properties, geometries, interactions and reactions. A particular issue which arises in the representation of chemical structures is the question of how much of this information to include. As each representation does not include all information about the molecule this transformation is not always unambiguous and unique (Gasteiger, 2003). The purpose of machine-readable structure representations is to mine the molecular information and to ensure that it is suitable for the most common operations on molecules such as storage/retrieval, identity, substructure/superstructure relationships, similarity and multivariate relationships (Bajorath, 2004; Leach & Gillet, 2007). In this section, the most widely used computer-readable representation of molecules are

described as well as a variety of methods to represent and compare molecules in chemical space.

2.2.1 Generation of computer molecular representations and descriptors from structure

2.2.1.1 Molecular Graph

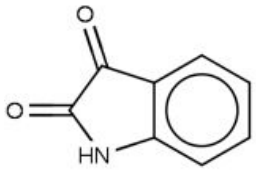
Molecular graphs serve as a convenient model for representing chemical structures in a computer. In a molecular graph (usually non-directed and connected multi-graph), the nodes correspond to the atoms and the edges to the bonds. Its vertices and edges are labelled with the kinds of the corresponding atoms or types of bonds, respectively. A graph represents only the topology of a molecule, that is, the way the nodes (or atoms) are connected and is less suitable for modelling those properties that are determined by molecular geometry, conformation or stereochemistry. The molecular graph can distinguish between structural isomers (compounds with the same molecular formula but non-isomorphic graph), furthermore, it normally does not contain any information about 3D arrangement and therefore cannot distinguish between conformational isomers or stereoisomers. Thus a given graph may be drawn in many different ways and may not obviously correspond to a "standard" chemical. The complexity of chemical systems is significantly reduced and some aspects are lost whenever they are modelled as graphs. It is necessary to have means to convert the molecular graph to and from a computer-readable format. This can be achieved in a variety of ways. Common methods are to use linear notations or graph matrices (Leach & Gillet, 2007). Connectivity of atoms through bonds leads to adjacency and distance matrices. The polynomials, generated from these matrices may be treated as the signature of those molecules. Several molecular descriptors are then calculated based on these polynomials (e.g. degree of a node, eigenvalues, distance-based topological indices, etc).

Linear Notations

Chemical line notations represent chemical structures as compact linear string of alphanumeric symbols, easily handled by computers and allowing fast manual

coding/decoding by trained users (faster than drawing a structure). Table 2.1 shows some examples of different line notations for the molecule 1H-Indole-2,3-dione.

Table 2.1: Different line notations for the molecule 1H-Indole-2,3-dione.



Systematic Name:	1H-Indole-2,3-dione
Name synonyms:	Isatin; Indole-2,3-dione; Isatic acid lactam; Isatine; 2,3-Diketoindoline; 2,3-Dioxo-2,3-dihydroindole; 2,3-Dioxoindoline; 2,3-Indolinedione; 2,3-Dihydro-1H-indole-2,3-dione; 3-Hydroxy-2-oxoindole
WLN:	T56 BMVVJ
ROSDAL:	1=2-3=4-5-6-1-9N-8-7-6,7=10O,8=11O
SLN:	O=C(N(H)C[2]=CC=CC=C(@4)@2)C[4]=O
SMILES:	O=C(C1=O)Nc2c1cccc2
Unique SMILES:	[NH]1[C]([C]([c]2[cH][cH][cH][cH][c]12)=[O])=[O]
InChI:	InChi=1/C8H5NO2/c10-7-5-3-1-2-4-6(5)9-8(7)11/h1-4H,(H,9,10,11)
InChIKey:	JXDYKVIHCLTXOP-UHFFFAOYSA-N

An early (1949) and remarkably compact fragment-based line notation that became quite widely used was the Wiswesser Line Notation (WLN) . This notation uses a complex set of rules to represent different functional groups (with more than 40 symbols) and the way they are connected in the molecule, which makes the notation difficult to code and error-prone. Most of the complexity of the notation is involved in determining the order in which the symbols are to occur so as to achieve not only a complete and unambiguous representation of the structure but also a unique or canonical representation (Wiswesser, 1954). From 1985 on, Representation of Organic Structures Description Arranged Linearly (ROSDAL) notation was developed by Welford, Barnard and Lynch granted by the Beilstein Institute. The ROSDAL generation process is straightforward, six rules

permit to code the organic molecule into an unambiguous but not unique alphanumeric string. Nevertheless, its use was restricted to Beilstein-DIALOG system (Barnard *et al.*, 1989). In 1988, Weininger (1988) at the US Environmental Research laboratory (USEPA) released the Simplified Molecular Input Line Entry System (SMILES) for chemical data processing, which has found widespread distribution as a universal standard chemical nomenclature. Compared with WLN and ROSDAL, SMILES is more intuitive and uses a set of six, very basic, rules to convert a structure into a string: (1) atoms are represented by their atomic symbols; (2) hydrogen atoms are omitted; (3) neighbouring atoms stand next to each other; (4) double and triple bonds are characterized by "=" and "#", respectively; (5) branches are represented by brackets; (6) rings are described allocating digits to the two connecting ring atoms (Weininger, 1988). Information about chirality and geometrical isomerism can also be included in the SMILES notation. The absolute stereochemistry at chiral atoms is indicated using the "@" symbol and geometrical (cis–trans) isomerism about double bonds is indicated using slashes. This notation has been later extended mainly by Daylight Chemical Information Systems Inc. and several coding enhancements derived from it such as SMiles ARbitrary Target Specification (SMARTS) for substructural pattern search and SMIles ReaKtion Specification (SMIRKS) for encoding reaction transformations (James *et al.*, 2011). A special extension of SMILES is Unique SMILES (USMILES), a canonical and unambiguous representation of a structure, granted by a proprietary algorithm (Weininger *et al.*, 1989). This has led to the use of different generation algorithms and/or different implementations, and thus, different SMILES of the same compound can be found. The SYBYL line notation (SLN) (1997) is a specification for unambiguously describing the structure of chemical molecules using short ASCII strings, and is a product of Tripos Inc. (Ash *et al.*, 1997). SLN was inspired by the SMILES notation but differs from it in several ways. SLN can specify molecules, molecular queries, and reactions in a single line notation whereas SMILES handles these through language extensions. SLN has support for relative stereochemistry, it can distinguish mixtures of enantiomers from pure molecules with pure but unresolved stereochemistry. In SMILES, aromaticity is considered to be a property of both atoms and bonds whereas in SLN it is only a property of bonds. The latest line notation is the IUPAC's International

Chemical Identifier (InChI) (Stein *et al.*, 2003). Like USMILES notation, the InChI allows a canonical serialization of molecular structure and unlike SMILES its main objective is to identify a compound in a unique and non proprietary manner. SMILES is certainly more human-readable by an expert and can be used for substructure search and analysis. Furthermore, InChI allows detection of tautomeric forms and group mobile hydrogen atoms together. Every InChI starts with the fragment "InChI=" followed by the version number. Structural information is organized in six layers and sub-layers, describing different aspects of a molecule. A special form of line notation as structure representation developed to facilitate searching is InChIKey. It is a condensed version created from InChI through hashing using the Secure Hash Algorithm InChIKey with a fixed length of 27 characters. InChI and InChIKey are currently used by several public and commercial databases as well as scientific journals.

Viewing, editing and converting chemical formats

Numerous computer applications are available to handle molecular structure information. Since chemical data has its own specificities, many formats were developed to facilitate information exchange (Gasteiger, 2003). The most widely used file formats in chemistry are summarized in Table 2.2.

When working with chemical information in cheminformatics, creating, querying, modifying and saving representations of chemical structures are very important tasks. For that propose, there are several molecule editors to manipulate chemical structure representations in either 2 or 3D. Typically, molecule editors support reading and writing at least one of the file formats mentioned above and they can mainly be divided into stand-alone programs and web-based applications (applets) (Gasteiger, 2003; Gunda, 2011). Table 2.3 summarizes the most widely used molecule editors/viewers. The wide variety of chemical structure representations in use has inevitably resulted in a need to interconvert them. OpenBabel (O'Boyle *et al.*, 2011, 2008) and JOELib (Guha *et al.*, 2006) are freely available open source tools specifically designed for converting between file formats.

Table 2.2: Summary of the most widely used file formats for exchange chemical structure information.

File Format	Description	More Information
Molecule-Data file – MDL Molfile (*.mol)	The MDL Molfile contains information about the atoms, bonds, connectivity and coordinates of a molecule. It is the most widely used connection table format.	http://accelrys.com/
Structure-Data file – SDfile (*.sdf)	The Structure-Data file is an extension of the MDL Molfile containing one or more compounds and the ability to include associated data.	http://accelrys.com/
Reaction-Data file – Rdfile (*.rdf)	The Reaction-Data file is an extension of the MDL Molfile containing one or more sets of reactions.	http://accelrys.com/
SMILES (*.smi)	SMILES is the most widely used linear text format which can describe the connectivity and chirality of a molecule.	http://www.daylight.com
Canonical SMILES (*.can)	The Canonical SMILES format (can) produces a canonical representation of the molecule in SMILES format.	http://www.daylight.com
Chemical Markup Language – CML (*.cml)	CML is an open standard for representing molecular and other chemical data. The open source project includes XML Schema, source code for parsing and working with CML data, and an active community. CML data files are accepted as input of many chemical applications.	http://www.xml-cml.org
IUPAC's InChI (*.inchi)	IUPAC's InChi file format which contains only structure definitions in a unique and predictable ASCII character string.	http://www.iupac.org/inchi/

2.2.1.2 Graph Tables and Matrices

Graph matrices effectively enable the molecular structure to be treated as a topological graph, i.e. as a set of atoms linked by bonds. This numerical description of the structure of chemical compounds is essential for computer manipulation of molecules and for calculation of various topological indices. Thus, a variety of graph matrices and tables have been proposed, such as atom connectivity table, adjacency matrix, bond matrix, 3D coordinate table and distance matrix (Todeschini & Consonni, 2009). Figure 2.3 shows some examples of different graph tables and matrices for the molecule 1H-Indole-2,3-dione. A connection table can have a high or low sophistication and can contain a large or small amount of information. The atom lookup table assigns arbitrarily a unique number to each atom, along with listing other properties such as its element type. At its most basic level the connection table represents which atoms are bonded to which other atoms, the bond order being indicated as an integer (i.e. 1 = single bond, 2 =

Table 2.3: Overview of the most widely used molecule editors/viewers and their details.

Computer Application	Platforms	Free	Description	More information
JChemPaint	All platforms/ web-based (applet)	Yes	JChemPaint is a 2D editor/viewer dealing with most of the formats described above.	http://sourceforge.net/apps/mediawiki/cdk/index.php?title=JChemPaint
ACD/ ChemSketch	Windows	Yes	ACD/ChemSketch is a 2D editor/ 2D and 3D viewer dealing with most of the formats described above.	http://www.acdlabs.com/products/draw_nom/draw/chemsketch/
Avogadro	Linux, Mac OS X, Windows	Yes	Avogadro is a 2D and 3D editor/viewer dealing with most of the formats described above.	http://avogadro.openmolecules.net/wiki/Main_Page
JME Molecular Editor	Web-based (applet)	Yes	JME Molecular Editor is a 2D editor/viewer dealing with most of the formats described above.	http://www.molinspiration.com/jme/index.html
Isis/Draw	Windows	Yes	Isis/Draw is a 2D editor/viewer dealing only with MDL molfile. It also allows to setup a database of structures.	http://accelrys.com/products/informatics/cheminformatics/draw/index.html
Accelrys Draw	Windows/ web-based (applet)	Yes/No	MarvinSketch is a 2D editor/ 3D viewer dealing with most of the formats described above.	http://accelrys.com/products/informatics/cheminformatics/draw/
MarvinSketch/ View	All platforms/ web-based (applet)	Yes/No	MarvinSketch is a 2D editor/viewer dealing with most of the formats described above.	http://www.chemaxon.com/products/marvin/
MarvinSpace	All platforms/ web-based (applet)	Yes/No	MarvinSpace is a 3D viewer dealing with most of the formats described above.	http://www.chemaxon.com/products/marvin/marvinspace/
ChemDraw	Windows, Mac OS X	No	ChemDraw is an advanced 2D editor/ 2D and 3D viewer dealing with most of the formats described above.	http://www.cambridgesoft.com/software/ChemDraw/
ChemDoodle	All platforms/web-based (applet)	Yes/No	ChemDoodle is a 2D editor/viewer dealing with most of the formats described above.	http://www.chemdoodle.com/

double bond, 3 = triple bond, additionally a 4 can be used for an aromatic bond). Normally, the hydrogen atoms are not shown because this atom usually establishes single connections to other atoms. Comparing with matrix representations the major advantage of connection tables is that the number of entries increases as a linear function of the number of atoms in the molecule instead of increasing with the square of the number of atoms (Gasteiger, 2003; Leach & Gillet, 2007;

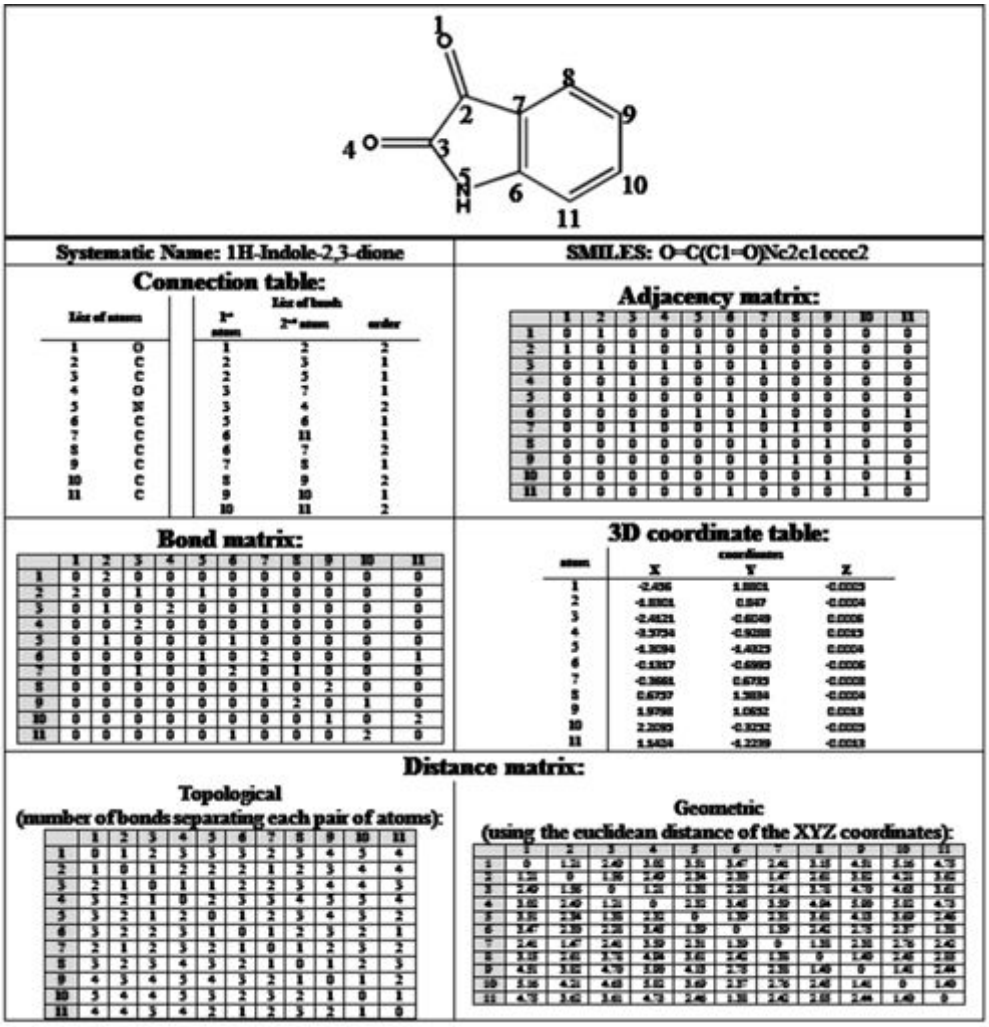


Figure 2.3: Different graph tables and matrices for the molecule 1H-Indole-2,3-dione.

Todeschini & Consonni, 2009). The adjacency matrix is a representation of a connection table however, no bond types or bond orders are presented; it contains only 0 and 1 (bits). The matrix is square and typically sparse, reflecting the adjacencies between atoms or bonds in the graph (Figure 2.3). The bond matrix is related to the adjacency matrix but gives information about the bond order of the connected atoms indicated as an integer (i.e. 1 = single bond, 2 = double bond, 3 = triple bond) (Figure 2.3) (Gasteiger, 2003; Leach & Gillet, 2007; Todeschini & Consonni, 2009). The 3D coordinate table XYZ Cartesian

coordinates of the atoms, allowing geometric distance calculation and generation of 3D view of chemical structures (Figure 2.3). The distance matrix contains values which specify the shortest distance between the atoms of a molecule. This distance can be expressed either as topological distances (based on the number of bonds between atoms) or as geometric distances (based on the XYZ coordinates of the atoms) (Figure 2.3) (Gasteiger, 2003; Todeschini & Consonni, 2009).

2.2.1.3 Molecular Descriptors

Methods based on structural descriptors attempt to describe the information encoded in the molecular structure into a set of numerical values and define some means for comparing them (Nikolova & Jaworska, 2003). Some molecular descriptors are derived from chemical structural representation models, while others are the chemical structural representation itself. *"The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment"* (Todeschini & Consonni, 2009). These molecular descriptors are numerical values and make possible the manipulation and analysis of chemical structural information. Thousands of different types of representations of chemical structures and molecular descriptors exist within the field of cheminformatics, differing in the complexity of the encoded information and in the computation time (Todeschini & Consonni, 2009). These descriptors can be classified based upon the degree of structural information required to compute them or their nature/applicability (Faulon & Bender, 2009; Gasteiger, 2003; Leach & Gillet, 2007; Todeschini & Consonni, 2009).

Computer-readable representation of molecules can be separated into four hierarchical types of molecular descriptors according to the dimensionality of the encoded information: **(1)** *one-dimensional* (1D) constitution-based representation which capture information that is slightly discriminative but fast to compute, including counts of atoms or summations of the properties of atoms present such as molecular formula, counts of atoms or bonds of each type and molecular weight; **(2)** *two-dimensional* (2D) topology-based representation including connectivity information with no explicit geometric information, such as linear

notations (for example, SMILES and InChI), connection tables, graph representations and counts of particular types of fragments (for example, fingerprints); **(3)** *three-dimensional* (3D) geometry-based representation including topological connectivity and geometric arrangement of atoms, such as the position of the atoms in space (XYZ coordinates), the angles and distances between the atoms in the molecule; **(4)** *four-dimensional* (4D) physico-chemical properties-based including molecular properties arising from interactions of the molecule with the surrounding space derived from computationally expensive empirical schemes or molecular orbital calculations, such as quantum mechanical calculations (e.g. dipole moment, partial charges and hyperpolarizability), flexibility of bonds and conformational behaviour. These categories of descriptors are heterogeneous and not always mutually exclusive, originating the appearance of different schemes of categorization in the literature (Gasteiger, 2003; Todeschini & Consonni, 2009).

In some studies it has been shown that on average, 2D descriptors perform as well or better than higher dimensional representations, saving computational effort to predict 3D structures or 4D molecular properties (Dixon & Merz, 2001; Matter & Patter, 1999).

Software for generation of molecular descriptors

There are many and diverse molecular descriptors and their generation from chemical structures constitutes a major problem in cheminformatics studies. Fortunately, there exists several free and commercial computer applications that automatically generate large sets of molecular descriptors (Gasteiger, 2003; Todeschini & Consonni, 2009). Each computer application has its advantages and disadvantages even though, in many cases different computer applications generate different values for the same molecular descriptor (Geronikaki *et al.*, 2008). Hence, selecting the computer application to generate each molecular descriptor is an important step of any cheminformatics study. A compilation of the most used computed applications for molecular descriptors calculation is presented in Table 2.4.

Table 2.4: Summary the most used software for molecular descriptors calculation.

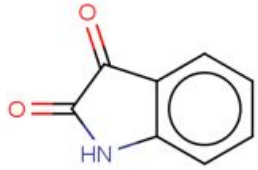
Computer Application	Platforms	Free	Description	More information
CDK[a]	Windows/Unix/MacOS	Yes	The CDK is a Java library for structural cheminformatics. CDK calculates over 200 molecular descriptors of several types.	http://sourceforge.net/projects/cdk/
E-DRAGON	Web-based	Yes	E-DRAGON is the electronic remote version of the software DRAGON, it provides more than 1,600 molecular descriptors that are divided into 20 logical blocks.	http://www.vcclab.org/lab/edragon/
PaDEL	Windows/Unix/MacOS	Yes	PaDEL calculates over 800 molecular descriptors (672 1D, 2D descriptors and 134 3D descriptors) of several types using CDK with some additional descriptors.	http://padel.nus.edu.sg/software/padeldescriptor/
MOLE db[b]	Web-based	Yes	The MOLE db is a free on-line database constituted of 1124 molecular descriptors calculated on 234773 molecules.	http://michem.disat.unimib.it/mole_db/
DRAGON	Windows/Unix	No	DRAGON calculates over 4800 molecular descriptors of several types: constitutional, topological, 2D-autocorrelations, geometrical, WHIM, GETAWAY, RDF, functional groups, properties, 2D binary and 2D frequency fingerprints, etc.	http://www.talete.mi.it/products/dragon_description.htm
CODESSA	Windows	No	CODESSA calculates over 1500 molecular descriptors of several types: constitutional, topological, geometrical, charge-related, semi-empirical, thermodynamical.	http://www.codessa-pro.com/
MOLCONN-Z	Windows/Unix	No	MOLCONN-Z calculates over 40 topological molecular descriptors.	http://www.edusoft-lc.com/molconn/
ADRIANA.code	Windows/Unix	No	ADRIANA.code calculates over 1240 molecular descriptors of several types: global physicochemical descriptors, atom property-weighted 2D- and 3D-autocorrelations and RDF, surface property-weighted autocorrelations.	http://www.molecular-networks.com/products/adrianacode
MOE	Windows/Unix/MacOS	No	MOE calculates over 600 molecular descriptors of several types: topological, physical properties, structural keys, etc.	http://www.chemcomp.com/software.htm
ADAPT[c]	Unix	No	ADAPT calculates over 260 molecular descriptors of several types: topological, geometrical, electronic and physicochemical.	http://research.chem.psu.edu/pcjgroup/adapt.html
ADMET	Windows	No	ADMET Predictor calculates 297 molecular descriptors (266 – 2D and 31 – 3D) of several types: constitutional, functional group counts, topological, E-state, Moriguchi descriptors, Meylan flags, molecular patterns, electronic properties, 3D descriptors, hydrogen bonding, acid-base ionization, empirical estimates of quantum descriptors.	http://www.simulations-plus.com/

[a]Chemistry Development Kit
[b]Molecular Descriptors Data Base
[c]Automated Data Analysis using Pattern Recognition Toolkit

2.2.1.4 Molecular Fragments

Molecular fragments are a particularly complex form of descriptors which are generally represented by fingerprints (Bajorath, 2004). Fingerprints are typically encoded as binary bit strings with different settings (such as generation method, length, size of patterns and number of bits activated by each pattern), representing a fragment or characteristic of a given molecule (James *et al.*, 2011). In other words, the fingerprint (the fixed-length bit-string) contains a fixed number of bits in which each bit can represent the absence (0) or presence (1) of some feature. Discrete variables, with more than two values, can be represented in the binary bit-string by using a bit for each possible value or for given ranges of values. Continuous variables can be represented by defining ranges of values and then assigning a bit to each range, a process known as binning (Gasteiger, 2003; James *et al.*, 2011). The ranges covered by each bin can be (1) separate or overlapping, (2) equidistant, (3) equifrequent, or (4) user-/application-defined . The fingerprints can be (1) directly-, (2) dictionary-, or (3) hash-based (James *et al.*, 2011). Descriptors with fixed limits can be directly assigned to positions in a fingerprint, with offsets being calculated to assign different groups or different descriptors to separate areas of the fingerprint. Dictionary or key-based bit-strings employ a dictionary that specifies correspondence between particular functional groups or fragments and bit positions in a fingerprint, with each entry (structural key) in the dictionary being assigned a bit position (screen number). Dictionaries of functional groups tend to be small, so all groups can be listed in the dictionary and assigned to a short bit-string. In some situations this can make the interpretation of the results of an analysis more straightforward, especially when relating specific fragments to activity. Rather than selecting a subset of fragments for inclusion in a dictionary, hash-based bit-strings are created by fitting all of the fragments into the bit-string (Table 2.5 shows an example of a hash-based fingerprint for the molecule 1H-Indole-2,3-dione). This can be achieved by hashing the fragment to generate one or more integers that fall within the length or a given sub range of the length, of the fingerprint. The more integers generated by the hash function, the more unique patterns can be superimposed on the bit-string, so the more fragments can be included (James *et al.*, 2011; Leach & Gillet, 2007). The

Table 2.5: Fingerprint (1024 bits) and bit set list for the example molecule 1H-Indole-2,3-dione.



Systematic Name:	1H-Indole-2,3-dione
SMILES:	O=C(C1=O)Nc2c1cccc2
Hashed Fingerprint:	00004022 00000800 00100100 00010600 20000000 00010010 00002000 00080000 04800800 000808c0 08280004 4000b000 01008002 00100000 40020140 00000000 00040000 82000000 08000000 00000000 40100000 18228600 20002000 81400015 00040000 80000280 00000010 80001080 04008000 00060003 80007000 20000600

<670>	C-C	<1006>	C4C4C4C-N-C-C	<652>	O=C-C-C
<260>	C-C-C	<710>	C4C4C4C4C	<45>	O=C-C-C4C
<609>	C-C-N-C-C	<223>	C4C4C4C4C-C	<201>	O=C-C-C4C-N
<693>	C-N-C	<699>	C4C4C4C4C-C-C	<632>	O=C-C-C4C4C
<9>	C-N-C-C	<111>	C4C4C4C4C-N-C	<574>	O=C-C-C4C4C4C
<948>	C-N-C-C-C	<655>	C4C4C4C4C4C	<653>	O=C-C-N
<140>	C-N-C4C-C	<912>	C4C4C4C4C4C-C	<46>	O=C-C-N-C
<479>	C-N-C4C-C	<711>	C4C4N-C-C	<44>	O=C-C-N-C4C
<81>	C4C	<711>	C4N-C-C-C	<199>	O=C-C-N-C4C4C
<473>	C4C-C	<82>	N-C	<762>	O=C-C=O
<723>	C4C-C-C	<906>	N-C-C	<63>	O=C-C4C
<341>	C4C-C-C-N-C	<348>	N-C-C-C	<848>	O=C-C4C-N
<159>	C4C-C-N-C	<711>	N-C-C-C-C	<759>	O=C-C4C-N-C
<135>	C4C-N-C	<122>	N-C-C-C4C	<258>	O=C-C4C4C
<372>	C4C-N-C-C	<280>	N-C-C-C4C4C	<971>	O=C-C4C4C4C
<893>	C4C-N-C-C-C	<287>	N-C-C-C4C4C4C	<337>	O=C-C4C4C4C4C
<256>	C4C-N-C-C-C4C	<335>	N-C-C4C4C	<330>	O=C-N
<905>	C4C4C	<317>	N-C4C	<64>	O=C-N-C
<936>	C4C4C-C	<561>	N-C4C-C	<691>	O=C-N-C4C
<65>	C4C4C-C-C	<29>	N-C4C-C-C	<164>	O=C-N-C4C-C
<9>	C4C4C-C-C-N-C	<993>	N-C4C4C	<787>	O=C-N-C4C-C=O
<498>	C4C4C-N-C	<242>	N-C4C4C-C	<596>	O=C-N-C4C4C
<317>	C4C4C-N-C-C	<674>	N-C4C4C4C	<301>	O=C-N-C4C4C4C
<561>	C4C4C-N-C-C-C	<382>	N-C4C4C4C4C	<836>	N-C-C-C-C
<347>	C4C4C4C	<550>	N-C4C4C4C4C4C	<552>	C-N-C-C-C
<278>	C4C4C4C-C	<813>	N4C4C-C-C	<997>	C4C-N-C-C
<10>	C4C4C4C-C-C	<623>	O=C	<715>	C4C4C4C4C4C
<443>	C4C4C4C-N-C	<329>	O=C-C	<747>	N-C-C-C4C

This is an example of a hash-based fingerprint generated by OpenBabel (O'Boyle *et al.*, 2011) and a description of the 79 bits on. As it can be seen, sometimes more than one fragment sets to the same bit (highlighted with grey color), in which case some information will be lost. ["C": carbon; "O": oxygen; "N": nitrogen; "-": single bond;"=": double bond; "4": aromatic bond].

same pattern always activates the same bit(s). Some of the advantages offered by structural keys over hashed fingerprints are the expressiveness of the features, the interpretability (one to one correspondence) and the possibility to consider the frequency of the patterns. On the other hand, hashed fingerprints consider a larger set of patterns; it is universal since there are no patterns to choose like in the structural keys and are easier to implement (Flower, 1998; James *et al.*, 2011; Wild & Blankey, 1999).

Several examples of methodologies for chemical binary representations generation exist, the most widely used are described below:

1. *Daylight Chemical Information Systems*[1] fingerprints are generated using a path- or hash-based approach. The molecular fingerprint is generated from a hash of all the unique connection paths (subgraphs) up to a maximum size (typically 7) into a fixed length bit string. Fingerprints may be folded to decrease the length and increase the bit density. Typical sizes for Daylight fingerprints are 512 or 1024 bits in length, but any power of two can be generated (James *et al.*, 2011).

2. *Molecular Design Limited (MDL)*[2] fingerprints are generated using a key- or dictionary-based approach. This type of fingerprint uses a pre-defined set of definitions and creates fingerprints based on pattern matching of the structure to the defined "key" set. This key based approach relies on the definitions to encapsulate the molecular descriptions *apriori* and does not "learn" the keys from the chemical dataset. MDL fingerprints could take on a maximum bit length of 966. No folding occurs with this type of fingerprint (Gasteiger, 2003; McGregor & Pallai, 1997).

3. *Barnard Chemical Information Systems (BCI)*[3] fingerprints are generated using a key or dictionary approach in which the keys for the fingerprint are first generated from the set and then implemented in the description. This combines the Daylight and MDL approaches. Typically the BCI dictionary

[1] http://www.daylight.com
[2] http://www.accelrys.com/
[3] http://www.bci.gb.com/

generates thousands of keys, resulting in molecular fingerprint bit lengths on the order of 5,000 bits (Gasteiger, 2003).

4. *Mesa Analytics & Computing, LLC (MACCS)*[1] fingerprints are generated using the 320 "drug-like" fragments published by MDL to generate 320 bit string representations as well as the 166 bit string representations based on MDL's original public dataset. The keys are generated from SMARTS pattern matching against the chemical dataset using the SMARTS matching algorithm in *OEChem* from *OpenEye Scientific Software* (Gasteiger, 2003; McGregor & Pallai, 1997).

Fingerprint representations are often used to search for similar molecules or substructures of a query compound since they provide a rapid and effective screening although with some false hits. The procedure requires the calculation of a fingerprint for the queried compound and then a search against the corresponding fingerprints of compounds in the database. For pair wise comparison of compounds, fingerprint overlap is determined as a measure of similarity and calculated using various coefficients (described below). Fingerprints have two main disadvantages: information loss, since it simply indicates the presence or absence of a given fragment rather than set bits for the number of matches and bit string saturation, because within big and complex molecules almost all bits are set so the overlapping is larger (Faulon & Bender, 2009; Flower, 1998; Gasteiger, 2003).

2.2.2 Molecular Similarity

Similarity, *like beauty*, is a ubiquitous, elusive and intuitive concept based on self-perspective (Rouvray, 1992). The similarity concept is rooted in science but has also been subject of study in philosophy. From the point of view of a philosopher a chemical "A" cannot be similar to a chemical "B" in absolute terms but only with respect to some measurable key feature "C" (Nikolova & Jaworska, 2003). On the other hand, from the point of view of a chemist, similar compounds should be judged by experts in terms of "approximately similar backbone

[1] http://www.mesaac.com/

and almost the same functional groups" (Nikolova & Jaworska, 2003). The problem lies in the degree of consistency between chemists, since they have different views on similarity and the results are biased and ambiguous (Lajiness *et al.*, 2004). Dmitri Mendeléev was one of the first using this concept of similarity for the formulation of the periodic table of elements in 1869 (Mendeleev, 1869). Through comparison of the similar atomic properties and chemical behaviour, Mendeléev was able to classify all the elements into a table leaving gaps for unknown substances. Defining molecular similarity basically consists of mapping "chemical space" (a representation of a molecule in structural or some property space) to one-dimensional space with entities of real numbers. Ideally similarity measures for molecules behave proportionally to all physical and biological properties of molecules in this representation. In other words, it groups together all molecules with similar physical and biological properties in a confined area of chemical property space. In practice, we are far away from reaching this goal, since molecular representations have to this day only been applied to specific problems of molecular similarity (Bender & Glen, 2004) and are dependent on the appropriate combination of chosen representations of molecules and on the metric used to quantify similarity between them (Bender *et al.*, 2003; Nikolova & Jaworska, 2003).

2.2.2.1 Similarity according to constitution

Constitutional similarity is defined in terms of the connectivity of each atom (representing a compound as a molecular graph) and is expressed by topological descriptors for the two dimensional features of a molecule (Nikolova & Jaworska, 2003).

2.2.2.2 Similarity according to configuration and conformation

The molecular similarity based on its configuration is defined in terms of the three-dimensional arrangement of atoms and characterized by the valence angles of all atoms linked to other atoms. Configuration is expressed by shape and volume descriptors and other descriptors accounting for the three-dimensional features of a molecule. The conformation similarity represents thermodynamically stable

spatial arrangement of its atoms. Calculating similarity values on basis of 3D-structures imposes the additional problem of conformational sampling and lacks two features: they are not independent of the size of a molecule and they do not describe additional properties. However, although computationally demanding, there are several reports which clearly demonstrate the advantage of considering the molecules as three-dimensional entities (Bajorath, 2004; Bender & Glen, 2004; Nikolova & Jaworska, 2003). Comparative molecular field analysis (CoMFA) is the most popular approach of analyzing 3D molecular similarity. It describes 3D structure-activity relationships in a quantitative manner by a collection of sampled points. Although the CoMFA fields are good descriptors for explaining activity, the problem with this approach is that it is difficult and time consuming to find comparable sampling point to align the molecules (Bender & Glen, 2004; Eckert & Bajorath, 2007; Nikolova & Jaworska, 2003).

It must be emphasized that many other types and variants of structural representation have been suggested for the computation of molecular similarity, including physico-chemical properties, quantum chemistry approaches, 3D pharmacophore patterns, multi-pole moments, etc. There is not an ideal methodology to measure molecular similarity and each one has its own strengths and weaknesses. However, the simple fingerprint continues to be the representation of choice for molecular similarity screening, not only because of its computational efficiency but also because of its demonstrated effectiveness in the many comparative studies that have been carried out. It is clear that the performance measurement of these approaches is critically dependent on the testing methods and molecules used in the dataset (Bajorath, 2004; Bender & Glen, 2004; Nikolova & Jaworska, 2003).

2.2.2.3 Similarity Metrics

Besides the molecular representation itself, the applicability of the similarity principle depends on the measure of (dis)similarity used to compare two such representations. Various (dis)similarity metrics exist that return a score indicating the level of similarity between molecules under comparison. Considering two molecules, A and B, represented using 2D binary fingerprints, where $\boldsymbol{a}$ is the

count of bits on/descriptors in molecule A but not in molecule B, $\boldsymbol{b}$ is the count of bits on/descriptors in molecule B but not in molecule A, $\boldsymbol{c}$ is the count of the bits on/descriptors in both molecules A and B and $\boldsymbol{d}$ is the count of the bits off/absent descriptors in both molecules A and B. Table 2.6 presents the most widely used metrics to compare chemical molecules based on fingerprints (Holliday *et al.*, 2002).

Table 2.6: The most widely used metrics to compare chemical molecules based on fingerprints.

Metric Name	Score Range	Formula
Cosine	[0, 1]	$\frac{c}{\sqrt{(a+c)*(b+c)}}$
Dice	[0, 1]	$\frac{2*c}{(a+c)+(b+c)}$
Euclid	[0,1]	$\sqrt{\frac{c+d}{a+b+c+d}}$
Forbes	[0,∞[	$\frac{c*(a+b+c+d)}{(a+c)*(b+c)}$
Hamman	[-1, 1]	$\frac{(c+d)-(a+b)}{a+b+c+d}$
Jaccard-Tanimoto	[0, 1]	$\frac{c}{a+b+c}$
Kulczynski	[0, 1]	$0.5 * (\frac{c}{a+c} + \frac{c}{b+c})$
Manhattan	[1, 0]	$\frac{a+b}{a+b+c+d}$
Matching	[0, 1]	$\frac{c+d}{a+b+c+d}$
Pearson	[-1, 1]	$\frac{(c*d)-(a*b)}{\sqrt{(a+c)*(b+c)*(a+d)*(b+d)}}$
Rogers-Tanimoto	[0, 1]	$\frac{c+d}{(a+b)+(a+b+c+d)}$
Russell-Rao	[0, 1]	$\frac{c}{a+b+c+d}$
Simpson	[0, 1]	$\frac{c}{min((a+c),(b+c))}$
Yule	[-1, 1]	$\frac{(c*d)-(a*b)}{(c*d)+(a*b)}$

$\boldsymbol{a}$ is the count of bits on/descriptors in molecule A but not in molecule B, $\boldsymbol{b}$ is the count of bits on/descriptors in molecule B but not in molecule A, $\boldsymbol{c}$ is the count of the bits on/descriptors in both molecules A and B and $\boldsymbol{d}$ is the count of the bits off/absent descriptors in both molecules A and B.

The main difference between Hamman, Pearson and Euclidean metrics, versus, the Jaccard-Tanimoto (from this point on simply referred as Tanimoto), Dice and Cosine metrics, is that the first effectively consider a common absence of attributes as evidence of similarity, whereas the latter do not. Also, Hamman and Euclidean metrics are useful only for comparisons of two molecules to the same

target but not for two independent pairs of molecules, for which Tanimoto metric is preferred. Additionally, Tanimoto metric has been the measure of choice for fragment-based chemical similarity and remains one of the most popular measures in chemical similarity (Holliday *et al.*, 2002; Nikolova & Jaworska, 2003). For that reason Tanimoto metric will be used in this work to compare similarities between the molecules. Figure 2.4 exemplifies the calculation of the molecular similarity/dissimilarity between three compounds based on the Tanimoto coefficient using binary fingerprints.

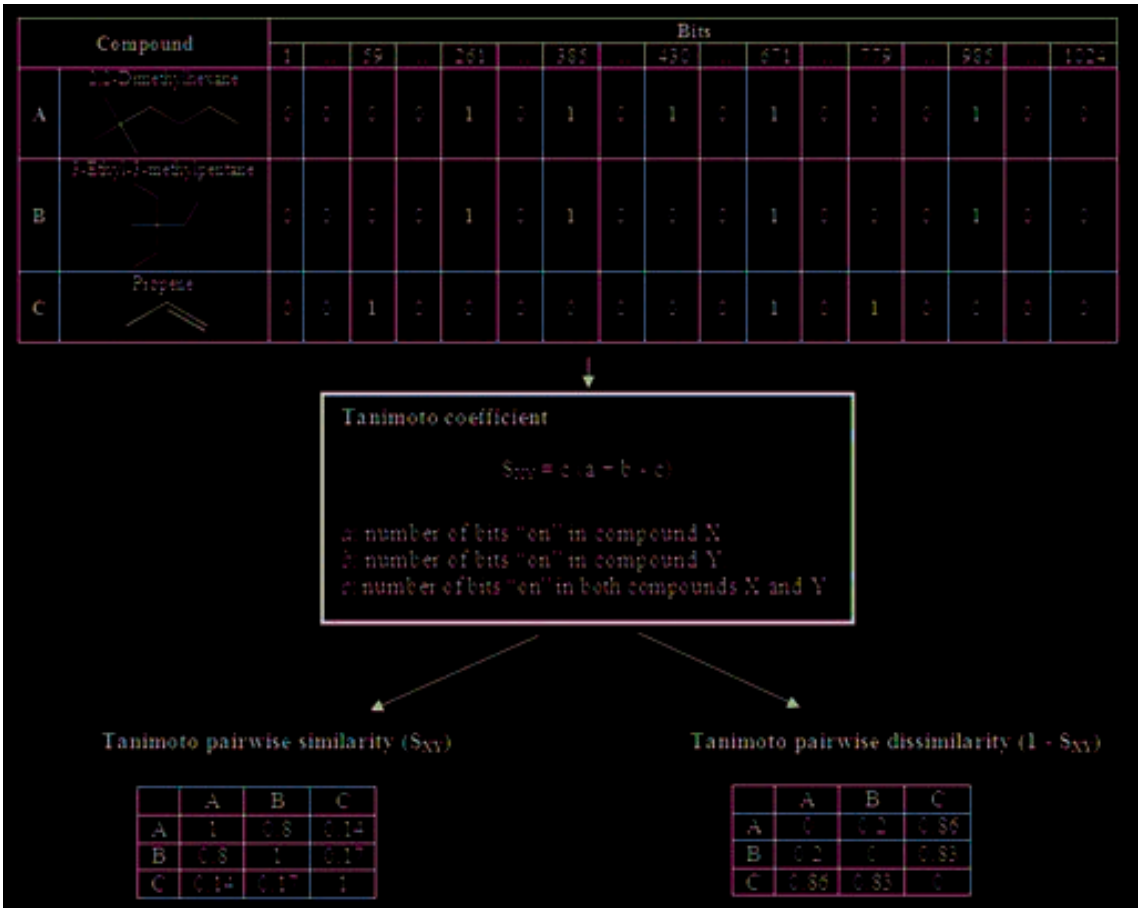


Figure 2.4: Schematic example of the calculation of the molecular similarity/dissimilarity between three compounds (A, B and C) based on the Tanimoto coefficient using binary fingerprints.

2.3 Approaches to select relevant molecular descriptors

"Men seek for vocabularies that are reflections of reality. To this end, they must develop vocabularies that are selections of reality. And any selection of reality must, in certain circumstances, function as a deflection of reality."

∼ Kenneth Burke, *A Grammar of Motives* (1945)

Molecular descriptors used in QSPR/QSAR are numerous and include examples discussed earlier in this chapter. Any set of descriptors may be used in QSPR/QSAR modelling, however only some of them are significantly correlated with each property in study. Furthermore, many of the descriptors are intercorrelated. This has negative effects on several aspects of QSPR/QSAR modelling, namely in terms of model overfiting, noise, computation time, extrapolation, predictive capacity and interpretability of the model (Dudek *et al.*, 2006). To tackle these problems, a wide range of methods for reducing the dimensionality or select the best combination of descriptors are used in QSPR/QSAR analysis. The first group of techniques, dimensionality reduction, aims to map the original high-dimensional data into a lower-dimensional space obtaining transformed features (linear combinations of the original features) while the second group of techniques, feature selection, aims to choose an optimal subset of features according to an objective function (Balakin, 2009; Bunin, 2007). The feature selection can be: **(1)** objective if it uses only the molecular descriptors (independent variables), removing redundancy amongst all the descriptors using the correlation matrix or **(2)** subjective if it also uses the property of interest (dependent variable) to identify the subset of descriptors that best map a relationship between structure and property using methods such as Genetic Algorithms or Simulated Annealing (Balakin, 2009). Clearly, it is impossible to give a detailed overview of all existing methods to reduce the dimensionality of a data set or select molecular descriptors but a general overview of the most widely used methods in QSPR/QSAR modelling is given below.

2.3.1 Dimensionality Reduction Methods

2.3.1.1 Principal Component Analysis

Principal Component Analysis (PCA) is a linear reduction method. It is a procedure based on the transformation of the descriptors space into linear orthogonal combinations that are ranked according to the explained variance of each combination of descriptors (named a principal component) (Han *et al.*, 2011). Thus, the first principal component is a linear combination of optimally-weighted observed descriptors that accounts for the maximal amount of total variance. The following components account for a maximal amount of variance in the observed descriptors that was not accounted for by the preceding components and they are linearly uncorrelated with all of the preceding components. PCA is fast to compute, easy to implement and several computer applications implement it (CRAN-R, 2012). This method guarantees to find a lower dimensional representation of the data on a linear subspace. The PCA method can only identify gross variability as opposed to distinguishing among and within groups' variability and the non linear combinations in the data cannot be efficiently exploited (Balakin, 2009; Cooley & Lohnes, 1971; Fodor, 2002). Principal components have been used as model inputs, when the variable space is too large and, specially, when models are particularly sensitive to the number of variables (e.g. Neural Networks) (Han *et al.*, 2011). Some examples of studies applying PCA for dimensionality reduction in QSPR/QSAR problems include Eriksson *et al.* (2006); Gramatica (2007); Katritzky *et al.* (2001).

2.3.1.2 Partial Least Squares

Partial Least Squares (PLS) is a linear reduction method. It performs linear combinations of molecular descriptors that maximize the explained variance in the dependent property by decomposing the input matrix of descriptors into two components, the scores and the loadings. The scores are orthogonal, being able to capture the descriptor information while reducing correlation between descriptors. The estimation of score vectors is done iteratively. The first one is derived using the first eigenvector of the property-descriptor combined variance-covariance matrix. Next, the descriptor matrix is deflated by subtracting the

information explained by the first vector. The resulting matrix is used in the derivation of the second score vector, which followed by consecutive deflation, closed the iteration loop. In each iteration step, the coefficient relating the score vector to the property is also determined (Cooley & Lohnes, 1971; Cramer, 1993; Dudek *et al.*, 2006; Fodor, 2002). PLS, in general, achieves better results than PCA when the dependent variable is known due to the supervised nature of its algorithm. PLS has a higher risk of overlooking correlations and sensitivity to the relative scaling of the molecular descriptors (Fodor, 2002). Some examples of studies applying PLS for dimensionality reduction in QSPR/QSAR problems include Eriksson *et al.* (2006); Goodarzi *et al.* (2013).

2.3.1.3 Multidimensional Scaling

Multidimensional Scaling (MDS) is a series of techniques that reduces the dimensionality by scaling the representation of the molecules based on a similarity or dissimilarity matrix into a reduced set of new variables (normally 2 or 3 dimensions). It performs linear and non-linear combinations of molecular descriptors that preserve distances between pairs or multiples of molecules with emphasis on recreating long distances. MDS is a very important technique to reduce dimensionality or to visualize a set of molecules described by means of the similarity or dissimilarity matrix. MDS has a substantial computational cost which makes this technique particularly crude or inapplicable to large data sets (Balakin, 2009; Cooley & Lohnes, 1971; Fodor, 2002). Some examples of studies applying MDS for dimensionality reduction in QSPR/QSAR problems include Gramatica *et al.* (2001); Sun *et al.* (2010).

2.3.1.4 Self-Organizing Map

Self-Organizing Map (SOM) is a linear and non-linear reduction method. It performs linear and non-linear combinations of molecular descriptors that preserve the linear and curved relationships between molecules in multidimensional space using a neighbourhood function emphasizing recreating short distances. On one hand SOM separates the molecules into a given number of clusters, on the other hand it visualizes these clusters on a 2D Kohonen map. The advantages of SOM

become especially important with increasing data dimensionality and size of the data set. SOM has a substantial computational cost to train the network (Cooley & Lohnes, 1971; Fodor, 2002; Kirew *et al.*, 1998). Some examples of studies applying SOM for dimensionality reduction in QSPR/QSAR problems include Maltarollo *et al.* (2013); Niculescu (2003).

2.3.2 Feature Selection Methods

2.3.2.1 Correlation- or Covariance-Based Methods

Correlation or covariance coefficients may serve as a preliminary filter for discarding intercorrelated descriptors. This can be done by for example, creating clusters of descriptors having correlation coefficients higher than certain threshold or retaining only one randomly chosen member of each cluster (Dudek *et al.*, 2006). Another procedure involves estimating correlations or covariances between a pair of descriptors and, if it exceeds or are below a threshold, randomly discarding one of the descriptors (Guha & Jurs, 2004). The choice of the ordering in which pairs are evaluated may lead to significantly different results. One popular method is to first rank the descriptors by using some criterion, and then iteratively browse the set starting from pairs containing the highest-ranking features (e.g. higher correlation between descriptors and property or higher covariance between descriptors). Normally this method is used as a preliminary step and in conjunction with other feature selection/reduction approaches.

2.3.2.2 Genetic Algorithms

A genetic algorithm (GA) (Goldberg, 1989; Goldberg & Holland, 1988) is a metaheuristic based on the application of a computational simplification of the biological evolutionary model over binary representations of solutions of a combinatorial optimization problem. Each solution is named a chromosome (or an individual), and its fitness is determined according to its result using an evaluation function. The algorithm starts by initiating a randomly generated set of solutions (named a population of chromosomes) and iteratively applies the evolutionary concepts of mutation, crossover and Darwinian selection to produce a new population. The process of selection is particularly important as an individual has a larger

probability of being selected for the new generation according to its fitness, leading each generation to become progressively better than the original one. The meta-heuristic process is repeated for a given number of iterations.

Genetic algorithms have been used for feature selection problems in QSPR and QSAR studies such as Bayram *et al.* (2004); Leardi (2001); Leardi & Lupianez Gonzalez (1998). For feature selection, generally a chromosome is modelled as a binary string identifying the selected features for a given prediction model. Typical models can be linear regression, Support Vector Machines or Neural Networks (Garrett *et al.*, 2003; Leardi & Lupianez Gonzalez, 1998; Ozdemir *et al.*, 2001; Tay & Cao, 2001). The evaluation function for each chromosome can then be a statistic of the application of the selected features using the predefined model to a validation set. Chromosomes with better validation results will tend to have a larger representation in the new population. The new population can then be changed using the crossover and mutation operators. Mutation changes randomly the solution by a fixed amount, causing some new features appear in the solution and others to disappear, therefore guaranteeing that all available features will have a chance of being evaluated during a set of generations. Crossover, on the other hand, will allow the exchange of features selected between chromosomes within the same generation. After mutation and crossover the new population is evaluated again and the process is repeated for a number of iterations or until a suitable solution has been found (Goldberg & Holland, 1988).

2.3.2.3 Simulated Annealing

Simulated Annealing (SA) is a stochastic method for the global optimization problem of locating a good approximation to the global optimum of a given function in a large search space by exchanging some percentage of the features in each iteration (Kirkpatrick *et al.*, 1983; Laarhoven & Aarts, 1987). SA works by emulating the annealing phenomenon from material science. Annealing is the physical process of heating up a material until it melts, followed by cooling it down in a controlled way until material crystallizes into a state with perfect lattice. During this process, the free energy of the material is minimized. The cooling process must proceed carefully in order to escape from locally optimal lattice

structures with crystal imperfections. This SA process for optimization can be formulated as a problem of finding a solution with minimal cost among the very large number of possible states. The physical annealing process can be modelled by computer simulation methods based on Monte Carlo techniques. The slower the cooling schedule, or rate of decrease, the more likely the algorithm is to find an optimal or near-optimal solution. Annealing with a slow cooling schedule is very slow and expensive computationally. The method cannot determine whether it has found an optimal solution (Dudek *et al.*, 2006; Fodor, 2002; Laarhoven & Aarts, 1987). Some examples of studies applying SA for feature selection in QSPR/QSAR problems include Ghosh & Bagchi (2009); Sharma *et al.* (2012).

2.4 Approaches to establish structure-property relationships using multivariate methods

"There are known unknowns; that is to say, there are things that we now know we don't know. But there are also unknown unknowns – there are things we do not know we don't know."
∼ Donald Rumsfeld, United States Secretary of Defense (2002)

Once the molecular descriptors are calculated and reduced to a subset of optimal descriptors the problem lies in building a model that better correlates the structure of the molecule with the desired property (Dudek *et al.*, 2006). A wide range of mapping function methods can be employed, including linear and non-linear ones. The linear models predict the property as a linear function of molecular descriptors and in general, they are easily interpretable and accurate for small datasets of similar compounds and molecular descriptors selected for the given property. The non-linear models predict the property as a non-linear function of molecular descriptors and in general, the models became more accurate, especially for large and diverse datasets but they are more complex and harder to interpret. Complex non-linear models may also fall prey to over-fitting (low generalization to unknown compounds during testing). In the framework of supervised learning another important division of the methods is based on the nature of the desired property: (1) classification tasks which approximate a discrete-valued function to map a pattern into a M-dimensional decision space,

where M is the number of categories or classes, and (2) regression tasks which approximate a real-valued target function to map a pattern into a continuous space. Furthermore, it is possible to follow two main strategies to predict properties of new compounds: (1) eager or model-based learning in which a model is build using a training set and then this model can be applied to all unseen cases to make predictions, and (2) lazy or instance-based learning in which each test instance is considered individually and information is extracted from the training set specifically for the prediction of that instance. The main advantage of lazy learning is that it is possible to make the most of the information about a test instance. It is impossible to give a detailed overview of all existing methods but a general overview of the most widely used methods in QSPR/QSAR is given below.

2.4.1 Instance-based Learning Approaches

Instance-based methods construct local approximations to the modelled function that applies in the neighbourhood of the new query instance. Thus it describes a complex target function as a collection of less complex local approximations based on the distance between instances. These algorithms have several advantages: they are simple but robust learning algorithms, can tolerate noise and irrelevant attributes, and can represent both probabilistic and overlapping concepts and naturally exploit inter-attributes relationships. Because the algorithm delays all processing until a new classification/prediction is required, significant processing is needed to make the prediction. Furthermore, the instances should be represented in such a way that allows the calculation of distance between them.

2.4.1.1 k-Nearest Neighbours

k-Nearest Neighbours (k-NN) is a simple method for classification or prediction that with increase in training data converges to the optimal prediction error (Itskowitz & Tropsha, 2005). The training phase of the algorithm consists only of storing the feature vectors and property values or classes of the training samples. For a given test compound, the method analyses its k-nearest neighbouring compounds from the training set and predicts the property based on the similarity

principle by *majority voting*, according to equation 2.1 where $N_k(x)$ is the neighbourhood defined by the k closest observations in the training set (Eklund *et al.*, 2014). The method is very sensitive to the metric used to map the compounds in the feature space and the training compounds available.

$$f(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i \tag{2.1}$$

2.4.2 Model-based Learning Approaches

Model-based approaches (Eklund *et al.*, 2014), on the other hand, represent what has been learned in a quantitative computational model that describes a mapping or transformation between a set of features and responses and that is richer than the language used to describe this data. Learning methods of this kind construct explicit generalizations of the training cases, rather than allowing generalization to flow implicitly from a similarity or distance measure.

2.4.2.1 Multiple Linear Regression

Multiple Linear Regression (MLR) models the property as a linear function of all the molecular descriptors weighted by coefficients adjusted and optimized from the training set (Dudek *et al.*, 2006; Kovdienko *et al.*, 2010). The coefficients are chosen to minimize the sum of square errors between the observed and predicted values of the property (Eklund *et al.*, 2014). This method is not appropriate to apply when handling a large number of descriptors per compound.

2.4.2.2 Partial Least Squares

Partial Least Squares (PLS) is a linear regression method that overcomes the MLR's problem of dealing with a large number of descriptors per compound (Dudek *et al.*, 2006; Wold *et al.*, 2001). The method assumes that the model is influenced by a relatively small number of latent independent variables. These linear combinations of the original variables are obtained as already explained in section 2.3.1.2 and are then used as input of a regression model. What distinguishes PLS from principal component regression is that, in PLS, the features

are weighted by the strength of their univariate effect on the output variable in the construction of each latent feature (Eklund *et al.*, 2014).

2.4.2.3 Artificial Neural Networks

Artificial Neural Networks (ANN) is a non-linear method for classification or prediction based on the parallel architecture of a biological neural network (Abraham, 2005; Dudek *et al.*, 2006; Eklund *et al.*, 2014). An ANN consists of a weighted interconnection group of artificial neurons that modulate the effect of the associated molecular descriptors represented by a transfer function. The learning capability of the ANN is achieved by adjusting the weights in accordance to the chosen learning algorithm. In supervised learning, an input vector of molecular descriptors is presented together with a set of desired property responses, one for each neuron, at the output layer. A forward step is done, and the discrepancies between the desired and actual property for each neuron in the output layer are found and used to determine weight changes in the net according to the learning rule.

2.4.2.4 Support Vector Machines

Support Vector Machines (SVMs) is a non-linear supervised learning algorithm used for a variety of classification and regression problems. Burbidge *et al.* (2001) published the first studies that featured SVMs tested in QSAR problems and this methodology proved superior to other machine learning tools, either in results or computational efficiency.

Differently from other methodologies based on heuristic optimization methods, SVMs are based on the solution of a convex quadratic programming problem, for which it is guaranteed to reach a minimum solution, which is deemed to be unique. The foundation of SVMs is the discovery of instances in the data (the support vectors) which construct a decision hyperplane or set of hyperplanes in a high-dimensional space that maximizes the margin according to a mathematical transformation of the variable space through a kernel function applied to the support vectors. Kernel functions are usually linear, polynomial, radial or

sigmoid, and generally machine learning libraries provide implementations to all these kernel functions.

Some of its unique characteristics include the capability of handling a very large number of descriptor variables with minimal over-fitting (as it is often the problem with other methodologies like ANN). The main disadvantages of SVMs are the lack of transparency of results due to its non-parametric nature and the sensitivity of the algorithm to the choice of kernel parameters (Burges, 1998; Dudek *et al.*, 2006).

2.4.2.5 Random Forests

Random Forests (RFs) are an *ensemble* method for classification or regression (Breiman, 2001). Ensemble methods are based on the iterative application of a simple classification or regression algorithm over a randomly defined subset of the data and use a *consensus* voting procedure for determining the outcome of its application. RFs use as a basic classification or regression algorithm, simple decision trees fitted where the leaves represent the property/activity value and branches represent conjunctions of descriptors that represent the structure of the compounds. Each tree is constructed independently of previous trees using a different bootstrap sample of data with replacement and where each node is split using the best subset of predictors randomly chosen at that node.

The basic process of RFs building can be summarily described in the following sequence of steps. The process is repeated once for every iteration ($i = 1..N$), according to a value specified by the user (N). One iteration will produce a simple decision tree from a set of variables and instances. For each fitted tree a distinct set of variables and instances is used. From the training dataset, a bootstrapping procedure is ran selecting with reposition a set of instances (Υ_i), with size equal to the training set. Also small subset of independent variables are specified by the user and randomly selected from all the available variables (Δ_i). Then a decision tree model $DT_i = f(\Upsilon_i, \Delta_i)$ is fitted to Υ_i and Δ_i. The set of all decision tree models (DT_i, where $i = 1, ..., N$) is a random forest. Using it for prediction implies running all trees to a new dataset and produce a *consensus* result from the classification or prediction outcomes of the individual decision trees. RFs

allow natively for an out-of-the bag validation, that is, each tree is validated with the instances that were not selected for its training (about one-third of the set) and global *consensus* statistic can be produced.

The generalization of this method depends on the strength of the individual trees in the forest and the correlation between them. The algorithm RFs has several characteristics that make it suitable for QSPR/QSAR datasets (Breiman, 2001; Statnikov & Aliferis, 2008): a) it can be used when there are more variables than observations; b) it has a good predictive performance even when noisy variables are present; c) it is not very sensitive to the algorithm parameters, therefore there is a minimal necessity to tune the default parameters to achieve a good performance; d) due to its nature encompassing a large number of simple models, it largely reduces the problems caused by over fitting; e) it can handle a mixture of categorical and continuous descriptors; f) it returns measures of descriptor importance; g) there are high quality and free implementations of the method. Furthermore, there is no need for cross-validation as it is estimated internally considering that each tree is constructed using a different bootstrap sample from the original data.

2.5 Approaches to assess and validate QSPR/ QSAR models

"The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data."
~ John Tukey (1915-200)

A validation of the developed models is an important aspect of any data-mining study. There are several procedures to assist the assessment and validation of a model and these can be used to determine whether the model is appropriate for the available data, to select the model with the best performance as well as to provide some estimate of how well the model can predict properties for new molecules, ensuring that the correlation is not due to chance factors and to avoid the possibility of overfitting (Puzyn *et al.*, 2009; Topliss & Edwards, 1979; Tropsha, 2006; Tropsha *et al.*, 2003). Most assessment statistics are different

for regression (predicting continuous data) and classification (predicting classes) models and are described below.

Suitable statistical validation of the models should be applied performing both internal and external validation. The internal validation assesses how the results will generalize to an independent data set using k-fold cross-validation (Gramatica, 2007; Tropsha, 2010). The external validation uses independent datasets that have not been used during the model development. These assessment statistics obtained from the results of the external validation give some indication of the predictive ability of the model. Another important component of the validation process is testing for chance correlations, i.e. test whether the results generated by the model were due to correlations rather than by a structure-property relationship. The simplest strategy to test for chance correlations is Y-randomization, this technique scrambles the dependent variable and calculates statistics of goodness-of-fit for the model using the scrambled dependent variable (Puzyn *et al.*, 2009; Tropsha, 2010).

2.5.1 Model fit

2.5.1.1 Regression models

The examination of the model fit is performed through the comparison of the experimental and predicted properties and is needed to statistically ensure that models are sound. The statistics coefficient of determination (R^2) and root-mean-squared error (RMSE) are widely used to determine the goodness of fit of regression models (Golbraikh & Tropsha, 2002; Tropsha, 2010). The use of RMSE shows the root of the error between the mean of the experimental values and predicted properties, corresponding to the mean of the squared error loss. The R^2 has a value between zero and one and indicates the proportion of the variation in the dependent variable that is explained by the model. Large RMSE values reflect the model's poor ability to accurately predict the properties even when a large R^2 value (≥ 0.75) is returned. The R^2 and RMSE are calculated according to the equations (2.2) and (2.3), respectively.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{2.2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n}} \tag{2.3}$$

where y_i is the experimental value of the property, $\widehat{y}_i$ is the estimated value of the property, $\overline{y}$ is the average property and n is the number of molecules in the set of data being examined.

2.5.1.2 Classification models

A classification model when applied to a given random molecule will provide an estimate of whether it belongs to one of two classes. To evaluate the results of a classification model, several commonly used statistics are extant, the most widely used are based on the contingency matrix, where the rows represent the classification according to the experimental data and the columns represent the predicted classes assigned by the model. The main diagonal represents the compounds correctly classified into each class, while the non-diagonal cells represent the misclassifications. The last column reports the number of compounds belonging to each class, whereas the last row reports the total number of compounds assigned to each class according to the model. The ability of a classification model to detect known positive instances (sensitivity or recall - equation (2.4)), known negative instances (specificity - equation (2.5)), positive instances against all compounds classified as positive (precision - equation (2.6)) and all chemicals in general (accuracy - equation (2.7)) indicate the overall performance of the model. The false positive (FP) and false negative (FN) rates can be calculated from the specificity and sensitivity, respectively. The true positive (TP) and true negative (TN) classification rates focus more on the effect of individual chemicals, since these are conditional probabilities. Thus, the positive classification rate is the probability that a chemical classified as active is really active, while the negative classification rate gives the probability that a chemical classified as inactive chemical is really inactive.

$$sensitivity\ or\ recall = \frac{TP}{TP + FN} \tag{2.4}$$

$$specificity = \frac{TN}{TN + FP} \tag{2.5}$$

$$precision = \frac{TP}{TP + FP} \tag{2.6}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.7}$$

These statistics on their own are unsuitable for unbalanced datasets as they will account similarly the contribution of each class. For instance in a classification problem with 2 classes A and B, where A contributes to 90% of the data, a trivial model that classifies everything as A will reach an accuracy of 90%. A more robust statistic is the *mean square contingency coefficient* (Baldi & Brunak, 2001) (also known as Matthews correlation coefficient), or ϕ:

$$\phi = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \tag{2.8}$$

This coefficient is more robust than simple accuracy to unbalanced datasets. The previous example will produce a ϕ of 0.0, which shows the actual inability of the model for classification, indistinct from random guessing. Another traditional measure of accuracy is the F-measure or balanced F1-score (equation 2.9) which is the harmonic mean of precision and recall, however contrarily to the mean square contingency coefficient, the F1-score does not take the true negative rate into account:

$$F1\ score = 2 * \frac{precision * recall}{precision + recall} \tag{2.9}$$

A good model has to score high in all described coefficients. The ϕ statistic provides a good overall metric, but to correctly identify a problem with the model it is also necessary to take into account sensitivity, specificity, accuracy and precision.

2.5.2 Internal cross-validation

Cross-validation is a common method of internally validating a model (Arlot & Celisse, 2010). The outcome from cross-validation is a cross-validated R^2 which is normally represented as q^2 (Consonni *et al.*, 2009; Golbraikh & Tropsha, 2002), RMSE, sensitivity, specificity, precision, accuracy and Matthews coefficient. The process of cross-validation begins with the random division of the dataset into n-folds of compounds. One partition is removed and used as test set and the model is created from the remaining data points. The process is repeated as many times as the number of folds in which the dataset was divided. The validation statistics are averaged over the rounds. The cross-validated statistics are used as criterion of robustness and predictive ability of the model and are calculated in the same way as described in the model fit statistics (Equations (2.2) to (2.8)) (Tropsha, 2010). These cross-validated statistics draw attention to the possibility of model overfitting. An overfitted model is an over generalized model for the training set and its ability to predict the properties of new molecules decreases. It is typical that non-cross-validated statistics are better than the cross-validated ones, yet overfitting of the model is usually suspected if the former values are significantly larger (typically $> 25\%$) than the latter (Andrew, 2001; Gramatica, 2007).

2.5.3 External Validation

An external validation is considered optimal when evaluating how well the model generalizes the new data (Consonni *et al.*, 2009; Gramatica *et al.*, 2007; Tropsha, 2010). The original data set should be randomly divided into two groups: the training set and the test set. The training set is used to derive a model that is further used to predict the properties of the test set members, which were not used in the model development. The repetition of random selection is a common practice for internal model validation, but it is not feasible for external validation because external objects cannot be reused to fit the model. This division is a problem if the size of the external test set is small, in such a situation new data should be collected.

2.5.4 Y-randomization

To establish model robustness, the Y-randomization (randomization of the dependent variable) test can be used (Tropsha, 2010). This test consists of repeating all the calculations with random scrambled properties of the training set. Ideally, calculations should be repeated at least five times. The goal of this procedure is to establish whether models built with real properties of the training set have good statistics not due to overfitting or chance correlation. If all models built with randomized properties of the training set have statistically significant lower predictive power (typically R^2 <0.5) for the training or the test set, then the models built with real properties of the training set are reliable (Bajorath, 2004; Faulon & Bender, 2009; Tropsha, 2010).

2.5.5 Applicability Domain

Formally, the developed model can predict the target property for any compound for which chemical descriptors can be calculated. However, it cannot be expected to extrapolate well for a compound considerably dissimilar to those used to develop the model (Tropsha, 2010). The expression *domain of applicability* of a model denotes the region of chemical space that is adequately represented by similar compounds in the training set, such that predictions within the domain will not suffer from this extrapolation problem (Faulon & Bender, 2009; Tropsha, 2010). There are several ways to determine the applicability domain, such as using distance- or similarity-based methods (e.g. Mahalanobis or Euclidian distance and Tanimoto similarity), coordinate-based range and probability density distribution-based methods (Netzeva *et al.*, 2005). The most used methods are based on chemical similarity between each test compound and K-Nearest Neighbours in the training set, but as stated before the assessment is not trivial, since the concept of similarity is subjective. The average similarity of the K-Nearest Neighbours is used for assessment of the applicability domain of the model. If the average similarity exceeds a selected threshold (typically 0.5) then the test chemical compound falls in the applicability domain of the model (Golbraikh & Tropsha, 2002; Netzeva *et al.*, 2005; Tropsha, 2010).

2.6 Summary

QSPR/QSAR are methods that allow the prediction of biological and physicochemical properties of molecules based on application of machine learning techniques to datasets of chemical compounds represented in machine-readable formats for which properties were measured experimentally. This chapter presented the underlying details of the fundamental steps in the development of a QSPR/QSAR: (1) quantifying the inherently abstract molecular structure, (2) determining which structural features most influence the given property (representation problem) and (3) establishing the functional relationship that best describes the relationship between these structure descriptors and the property/activity data (mapping problem) using machine learning algorithms. Such mathematical representation of the knowledge is referred to as a predictive model, which aims to predict the investigated property for new compounds. In this chapter, approaches to assess and validate these predictive models were also described. The field of property prediction based on the quantitative relationship between structure and property is certainly much more detailed than it has been described in this chapter and, the literature describes numerous algorithms and variants that are suited for both general use as well as for special cases. However, the focus of this chapter is to give a general overview of the most widely used approaches to overcome each step of a QSPR/QSAR study.

Chapter 3

Model-based Methods for Quantitative Structure-Property Relationship Modelling

There is a growing interest in the application of data-mining techniques in chemistry, mainly due to their flexibility in modelling non-linear relationships. Data mining is the analysis of (often large) data sets to find relationships and to summarize it in novel ways that are both understandable and useful to the data owner (Hand *et al.*, 2001). A molecule can have arbitrary dimension, structure and composition, and it is known that similar compounds many times share physical/chemical properties and biological activity. However, there is not a uni-vocal and unequivocal way of coding and comparing these molecules. Consequently, data-mining research still faces difficulties when handling this kind of non-homogeneous data that is not easily distinguished through a finite set of attributes. The literature describes two possible multivariate approaches to deal with non-homogeneous data: a) model-based and b) instance-based approaches. This chapter is centred on the former, decomposing the information available in a set of attributes and use standard data-mining/statistics techniques such as Multivariate Regression, Neural Networks, Support Vector Machines, Random Forests and Bayesian Networks (Falcao *et al.*, 2006; Han *et al.*, 2011; Young, 2009). This approach is normally used to develop QSPR/QSAR models or additivity schemes. The process of model development is typically and generally

divided into three steps: data preparation, data analysis, and model validation.

The aim of this work is to improve the current models for the prediction of properties of chemical molecules using advanced automated analysis solutions based on data-mining techniques. For this purpose, the study started with the exploration of different ways to represent chemical data and to transfer the various types of representation into a machine-readable representation and manipulation. Furthermore, one of the main topics in the development of QSPR/QSAR predictive models is the identification of the subset of variables that represent the structure of a molecule and which are predictors for a given property. There are several automated feature selection methods, ranging from backward, forward or stepwise procedures, to further elaborated methodologies such as evolutionary programming. The problem lies in selecting the minimum subset of descriptors that can predict a certain property with a good performance, computationally efficient and in a more robust way, since the presence of irrelevant or redundant features can cause poor generalization capacity. In this chapter an alternative selection method, based on Random Forests to determine the variable importance is proposed in the context of QSPR/QSAR regression and classification problems and compared with models that do not apply a feature selection step and with other feature selection methodologies widely used in other studies. The subsequent predictive models are trained with SVMs introducing the variables sequentially from a ranked list based on the variable importance. This chapter is structured in several steps for QSPR/QSAR development for different case-studies (presented in Appendix A): dataset preparation, model development, model validation and evaluation and the blending of the results for each case-study and discussion.

3.1 Related Work in QSPR/QSAR Modelling

Prediction is very difficult, especially about the future.
∼ Niels Bohr

An area where data-mining techniques are increasingly playing an important role is cheminformatics, considering that the number of known and synthesized chemical compounds is growing exponentially, but the determination of their properties as well as biological activities is a time consuming and costly process

and is lagging severely behind (Chen, 2006; Gasteiger, 2003). These complex non-homogeneous data lead to the development and application of data-mining tools to extract and understand the underlying QSPR/QSAR (Doucet & Panaye, 2011; Katritzky *et al.*, 2000, 2002). QSPR/QSAR methods, as described in Chapter 2, are widely used for prediction and their goal is to relate molecular descriptors, from molecular structure, with experimental chemical, physical and/or biological properties by means of data-mining methods (Katritzky *et al.*, 1995, 1997; Puzyn *et al.*, 2009; Tropsha, 2010; Tropsha & Golbraikh, 2007). The three major difficulties in the development of QSPR/QSAR models are **(1)** quantifying the inherently abstract molecular structure, **(2)** determining which structural features most influence the given property (representation problem) and **(3)** establishing the functional relationship that best describes the relationship between these structure descriptors and the property/activity data (mapping problem) (Dearden *et al.*, 2009; Puzyn *et al.*, 2009; Tropsha, 2010; Tropsha & Golbraikh, 2007; Yasri & Hartsough, 2001). The first difficulty can be overcome by the use of calculated molecular descriptors, developed to quantify various aspects of molecular structure (Karelson, 2000). In fact, this approach is one of the causes of the second difficulty since thousands of molecular descriptors are currently extant (Karelson, 2000; Todeschini *et al.*, 2008). The problem lies then in the identification of the appropriate set of descriptors that allow the desired property of the compound to be adequately predicted. To accomplish this and to find the optimal subset of descriptors that describes the relationship between the structure and the property/activity data, several statistical and data-mining methods are commonly used for dimensionality reduction and feature selection (Dutta *et al.*, 2007; Liu, 2004). Frequently, these descriptors are also selected based on expert knowledge about the problem, for example, the molecular weight of a drug is known to be an important parameter that may affect the capacity of a drug to permeate across the blood-brain barrier (Pardridge, 2005). However, in general, this task cannot be completely achieved manually, given the complex non-linear nature of the structure-property/activity relationships and the high number of existing molecular descriptors. An optimal solution for this problem requires an exhaustive search over all possible subsets. Considering the high number of

molecular descriptors (n) and the limited knowledge on the amount of necessary descriptors (p), it is required to try for each p the sum of the n^{th} row of the binomial coefficients, which involves 2^n possible combinations. This exhaustive enumeration of each subset is computationally impractical, except for small problems. Therefore, a reasonable alternative is then the use of heuristic approximations that may be able to find the best possible subset of descriptors within the available computational resources (Blum & Langley, 1997).

Several studies have investigated approaches to solve the descriptor selection problem in QSPR/QSAR (Dehmer *et al.*, 2012; Gonzalez *et al.*, 2008; Liu & Long, 2009). Any set of descriptors may be used in a QSPR/QSAR model and therefore techniques to reduce the dimensionality or select the best combination of descriptors are very important (Dehmer *et al.*, 2012). The first group of techniques, dimensionality reduction, aims to map the original high-dimensional data into a lower-dimensional space obtaining transformed features (generally linear combinations of the original features) (Dehmer *et al.*, 2012). The construction of models based on dimensionality reduction such as principal component analysis (PCA) (Xue *et al.*, 1999) and partial least squares regression (PLS) (Roy & Roy, 2008) compress the original dataset generating a smaller number of variables. PCA transforms the original dataset into orthogonal components, constructed by linear combinations of the existing variables. These are arranged in descending order according to the percentage of variance each component explains. Therefore the first components (principal components) are expected to translate the main sources of variability of the data, and may be better suited for modelling purposes (Dehmer *et al.*, 2012). However, PCA does not reduce the number of features needed for prediction, it only reduces the number of parameters in the model, as all features may be present in each component. The second group of techniques, feature selection, aims to choose an optimal subset of features according to an objective function (Balakin, 2009; Dehmer *et al.*, 2012). The feature selection can be: **(1)** objective if it uses only molecular descriptors (independent variables), removing redundancy amongst all descriptors using the correlation matrix or **(2)** subjective if it also uses the property of interest (dependent variable) to identify the subset of descriptors that best map a relationship between structure and property (Mosier & Jurs, 2002). For that purpose, several search algorithms have

been devised, ranging from simple heuristic approaches (Frohlich *et al.*, 2004; Xu & Zhang, 2001) which perform a *"greedy"* search of the best subsets of variables such as forward selection, backward elimination or stepwise procedures to further elaborate methodologies including simulated annealing (Sutter *et al.*, 1995) and evolutionary programming (Kubiny, 1994) such as genetic algorithms (Cho & Hermsmeier, 2002). These methods allow a stochastic evolutionary search of the possible solution space of a problem aiming for the selection of an optimal non-redundant set of variables, if sufficient computational resources are provided (Dehmer *et al.*, 2012). Other recent articles present multi-phase methodologies, in which the subsets of descriptors are selected and assessed using different algorithms (Soto *et al.*, 2009). The problem lies in selecting the minimum subset of descriptors that can predict a certain property with a good performance, less computational/time cost and in a more robust way, since the presence of irrelevant or redundant features can cause a poor generalization capacity.

3.2 Feature Selection using variable importance

In this chapter, we present an alternative approach to select molecular descriptors inspired by a methodology proposed by Genuer *et al.* (2010) and applied to different case-studies and molecular descriptors described in Appendix A, namely **(1)** case A - predicting thermochemical properties, **(2)** case D - NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge and **(3)** case G - Blood-Brain Barrier (BBB) penetration modelling. Genuer *et al.* (2010) proposes a two-steps procedure: (1) preliminary elimination and ranking, sorting the variables in decreasing order of standard deviation of Random Forests scores of importance from a series of runs and elimination of variables with small importance; (2) variable selection for prediction, starting from the ordered variables by constructing an ascending sequence of Random Forest models, testing the variables stepwise and retaining it only if the error gain exceeds a certain threshold. The algorithm Random Forest is widely used in the prediction context (classification and regression) given that it has several features that make it suitable for a QSPR/QSAR dataset (Breiman, 2001; Genuer *et al.*, 2008; Statnikov & Aliferis, 2008). These include good predictive performance even when there are more variables than observations, capacity

to handle a mixture of categorical and continuous descriptors, measures of descriptor importance and due to its nature encompassing a large number of simple models, it largely reduces the problems caused by over fitting. However, there are few works in the literature using Random Forests in the context of descriptor selection. To the best of our knowledge, beyond the work of Genuer *et al.* (2010), there is another study in the literature that uses Random Forests for gene selection in classification problems (Diaz-Uriarte & Alvarez de Andres, 2006), for that purpose several forests are generated iteratively and at each iteration the variables with the smallest variable importance are discarded; the selected set of variables is the one that yields the smallest prediction error. In this study we propose a hybrid approach that also uses Random Forests, but differently from Genuer *et al.* (2010), using the quantification of the average variable importance from a series of runs provided by this method, as a tool for molecular descriptors selection. This ranking can be used to build a predictive model, without eliminating any variables, using any other machine learning prediction method, in this case and differently from Genuer *et al.* (2010), Support Vector Machines (Cortes & Vapnik, 1995), inserting the variables stepwise in order to find a good balance between the number of variables and prediction error.

The two main objectives of this hybrid methodology are: (1) obtain a set of descriptors that are most related to the property of interest using the variable importance index calculated by Random Forests and (2) obtain the smallest possible set of molecular descriptors that can still achieve a good predictive performance that generalizes well even if the ratio between the number of variables and number of observations becomes unfavourable. In order to assess results, and have a reference of the developed models performance, for each case-study the results will be compared with the ones obtained for models without a feature selection step and for models using other feature selection or dimensionality reduction techniques such as Principal Components Analysis and Genetic Algorithms. Finally, the model performance will be tested using test or independent validation sets, the prediction error and selected molecular descriptors will be analysed and discussed.

3.3 Data and Methods

We never do anything well till we cease to think about the manner of doing it.
~ William Hazlitt

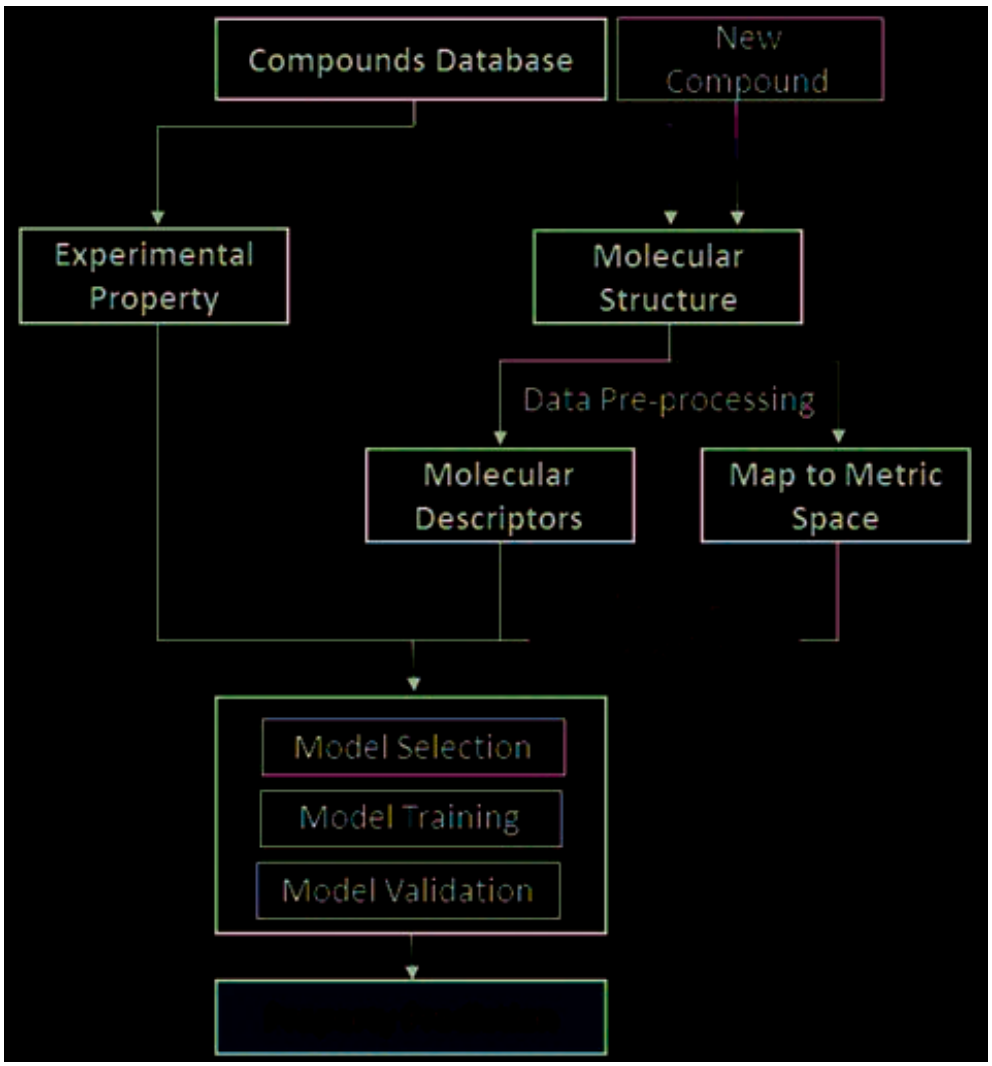


Figure 3.1: Flow chart showing the general steps used to predict properties in the context of QSPR/QSAR problems using model-based approaches.

The process of model development in QSPR/QSAR is generally divided into three steps (Figure 3.1): data preparation, data analysis, and model validation (Gramatica, 2007; Puzyn *et al.*, 2009; Tropsha, 2010; Tropsha & Golbraikh, 2007;

Yasri & Hartsough, 2001). The first stage includes the collection and cleaning of a dataset for the study and the selection of the best molecular representations. The second stage deals with the selection of a statistical multivariate data analysis and correlation techniques. The third stage validates and evaluates the developed model. As the problem discussed in this study is centred on models for feature selection and selection of statistical multivariate data analysis and correlation techniques, the second stage was performed several times using different sets of molecular descriptors as the purpose was to iteratively search for the optimal parameters for a model or for establishing the minimal number of variables necessary for adequately fitting a model without losing its predictive power for different case-studies. In order to ensure minimal bias in evaluating the results an exhaustive validation procedure was followed, both for model selection as well as for final model assessment. Therefore, during the model evaluation phase, each model was always internally validated using ten-fold cross validation or out-of-bag validation. After selecting a final model with a predefined set of variables, it was further validated with an external/test validation set never used on any phase of the training process and descriptor selection.

For the present section, initially the case-studies and molecular descriptor sets described in Appendix A used to validate this methodology are referred, followed by the main modelling methodologies applied. Also described are the procedures used for dimensionality reduction or feature selection either based on Random Forests variable ranking, principal components analysis and genetic algorithms.

3.3.1 Data and data pre-processing

As mentioned above, the case-studies and molecular descriptors used to validate this methodology are thoughtfully described in Appendix A. This research is based on the assumption that there is an underlying relationship between molecular structure and properties. Also, it is assumed that the multivariate molecular representation of the set of compounds reveals these analogies, i.e. physical, chemical and biological properties of chemical substance can be computed from its molecular structure, encoded in a numerical form with the aid of various descriptors. The key step in developing models is the selection of an informative and

representative dataset. The case-studies that were used to validate this methodology are : (1) case A (A1 and A2) - predicting thermochemical properties, (2) case D - NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge and (3) case G - Blood-Brain Barrier (BBB) penetration modelling. In addition to the descriptor sets that were already presented, other descriptor sets were also studied, however we present only those which obtained better predictive results. Also different combinations between the molecular representations were tried. One of the aspects that is considered important is the ability to interpret the models in a physico-chemical or biological sense. Thus, in the major part of the parameter sets the selection of descriptors was limited to those that seem to carry some fundamental physico-chemical or biological information that might be related to the modelled property.

3.3.2 Modelling Methodology

The second step pertains to creating a model that represents the relationship between structure and property. The sets of molecular descriptors serve as independent variables and the modelled property as the dependent variable. Many different data-mining methods are available and some of them have been described in Chapter 2. In this study four widely used techniques were used, a multivariate statistical method for reducing dimensionality of data: Principal Components Analysis (PCA), a method to select features: Genetic Algorithms (GAs) and two non-linear methods to build a predictive model: Support Vector Machines (SVM) and Random Forests (RF). The implementation of the methods is provided by R software in the packages princomp (PCA), e1071 (SVM) and randomForest (RF). In this section these methodologies will be briefly described, focusing on the implementation and parametrization details. The new hybrid approach using RFs for feature selection will be described in greater detail, as it is one important contribution of this work.

3.3.2.1 Approaches for model generation

Support Vector Machines (SVMs)

SVMs (Cortes & Vapnik, 1995) are non-linear supervised learning methods for classification or prediction. This algorithm can optimize the function to a global optimum and the results have good predictive performance (Burges, 1998; Dudek *et al.*, 2006), being currently one of the most used methodologies for QSPR/QSAR studies (Vyas *et al.*, 2014). The disadvantage of SVMs is the lack of transparency of results due to its non-parametric nature and the sensitivity of the algorithm to the choice of kernel parameters. It produces good results and generalizes well even if the ratio between the number of variables and the number of observations becomes very unfavourable or in the presence of highly correlated predictors. Another advantage is the kernel-based system since it is possible to construct a non-linear model without explicitly having to produce new descriptors. The accuracy of an SVM model is dependent on the selection of the model parameters. An Epsilon-Support Vector Regression analysis using the Gaussian radial basis function (RBF) kernel (general-purpose kernel used when there is no prior knowledge about the data) has two parameters: cost (represents the penalty associated with large errors, increasing this value causes closer fitting to the training data) and gamma (controls the shape of the separating hyper plane, increasing this value usually increases the number of support vectors).

For the present study, the SVM implementation used was provided by the e1071 (Meyer *et al.*, 2012) package from R. This library provides an interface to libsvm which allows classification or regression (Chih-Chung & Chih-Jen, 2001; Karatzoglou *et al.*, 2006). Using Epilson-SVMs with a preliminary tuning of the radial basis function (RBF) kernel parameters (which included the cost parameter that controls the trade off between allowing training errors and forcing rigid margins with the value 100 and the gamma parameter that controls the shape of the separating hyper plane with values ranging from 1×10^{-3} to 1×10^{-6} depending on the number and nature of descriptors used). Hyperparameter tuning in SVM models is done using the tune framework which is computationally expensive, considering that it performs a grid search over cost and gamma ranges.

Random Forests (RFs)

RFs, as already described, are a non-linear consensus method for classification or regression that ensemble unpruned decision trees for a good generalization. RFs have two model parameters that condition the model results, namely, the number of variables randomly sampled at each node to be considered for splitting and the number of trees in the forest. A preliminary systematic evaluation of both parameters on the training set was performed for each model. In RFs, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error. It is estimated internally considering that each tree is constructed using a different bootstrap sample from the original data. About one-third of the cases are left out of the bootstrap sample (out of the bag (OOB) samples) and not used in the construction of the forest. These OOB samples are used to get a running unbiased estimate of the regression error as trees are added to the forest and they are also used to get estimates of variable importance. The proportion of variation explained indicates how well the set of molecular descriptors is able to explain the variation in the property/activity value.

The Random Forest implementation used in this work was provided by the R library randomForest (Liaw & Wiener, 2002).

Variable importance

The ensemble voting procedure of RFs allows for the calculation of an importance score for each variable in the model. There are several available measures of variable importance. One of the most common measures is determined by looking at how much prediction error increases when the value of a variable in a node of a tree is permuted randomly while all others are left unchanged according to equation 3.1 (Biau, 2012; Breiman, 2001; Genuer *et al.*, 2008, 2010).

$$VI(X^j) = \frac{1}{ntree} \sum_t (\widetilde{errOOB_t^j} - \widetilde{errOOB_t}) \tag{3.1}$$

However, there is an issue in determining the variable importance of correlated variables, considering that in this determination it is assumed that each variable is independent of the response variable as well as from all other predictors (Strobl *et al.*, 2008). Therefore, if correlated predictors are not independent, they obtain high importance scores and consequently, a higher probability of being selected

for the split. Nevertheless, some correlated variables do hold predictive value, but only because of the truly important variable (Strobl *et al.*, 2008).

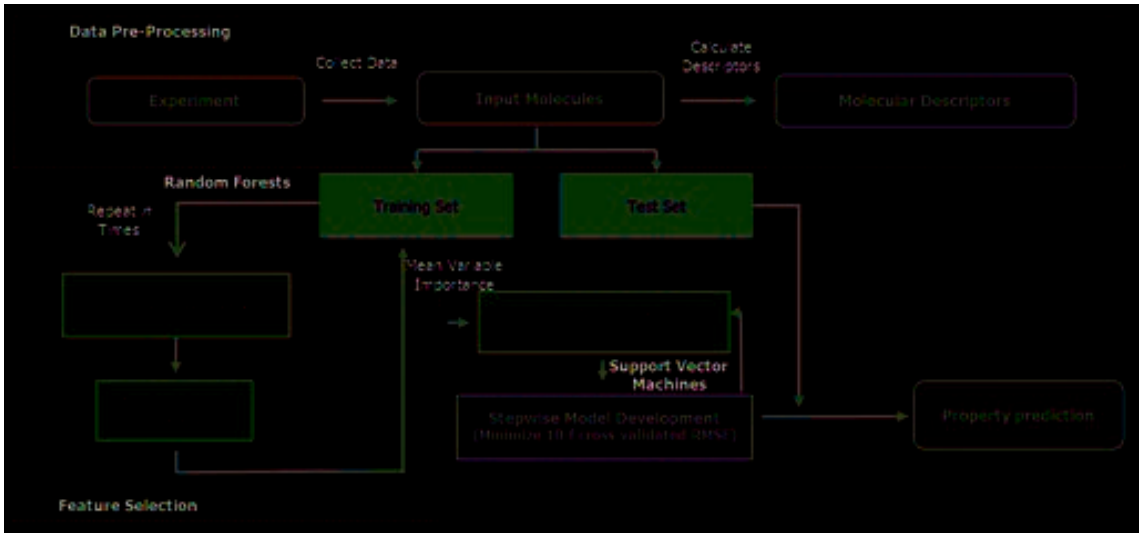


Figure 3.2: General workflow of the hybrid approach using Random Forests for Feature Selection for predicting properties of compounds based on molecular descriptors.

3.3.2.2 Approaches for feature selection

Variable importance for feature selection

It is possible to use the variable rankings according to their importance in RFs models as a criterion for variable selection in other models (Genuer *et al.*, 2008, 2010). The procedure followed in this work involved a sequence of steps in order to ensure coherence and results reproducibility (Figure 3.2). Therefore the procedure can be schematized as in Figure 3.2 with the following sequence of steps: **(1)** For the problem in study, an initial systematic evaluation of the optimal model parameters was performed, and the results with the out-of-bag (OOB) root mean square error were evaluated for selecting the best possible parameter combination; **(2)** With the best parameter set, perform n (it was found that $n > 10$ does not show any advantage) model runs and record each variable importance score and rank, and using this new consensus ranking, define

a sorted list starting with the most relevant variables and ending with the less important ones; **(3)** Proceed stepwise by feeding another prediction model (as an SVM) a progressively larger vector of input variables, following the ranked order. With such procedure it is expected that a minimal descriptor set, significantly smaller than the initial variable list may be found.

Genetic algorithm (GA)

As previously described, GA is a meta-heuristic based on the application of a computational simplification of the biological evolutionary model over binary representations of solutions of a combinatorial optimization problem.

A GA was adapted to this problem and implemented considering the following parameters: a) the number of chromosomes – this parameter indicates how many solutions are being evolved simultaneously; b) the mutation rate – indicates the likelihood of a given feature is swapped into or out from a solution (chromosome) a value of 0.05 indicates that each feature has a probability of 5% of being changed. To avoid large drifts, the only mutation possibility is a swap, meaning that for each feature that leaves the solution, another one, not previously there, has to enter; c) the crossover rate – indicates how likely two chromosomes can exchange variables in the models; d) the solution density – indicates how many features can be selected for each solution; e) the selection factor – indicates the likelihood that a given solution can be selected for the new population and it is a function of its rank among the current population, better chromosomes mean that the respective solution or combination of features produces an improved model compared to the others. Superior models are ranked higher, and higher ranking models have an increasingly large probability of selection using a negative exponential distribution. Smaller values of the selection factor indicate a very small probability of selecting the worst solutions for the new generation, while larger values emphasize the possibility of selecting substandard solutions. All parameters were subject to a preliminary optimization process, so that it was possible that the implementation could explore a significant fraction of the solution space.

Principal component analysis (PCA)

PCA, as already described, is a procedure based on the transformation of the variable space into linear orthogonal combinations that are ranked according to the explained variance of each combination (named a principal component).

The procedure followed involved a sequence of steps in order to ensure coherence and reproducibility of results. This procedure can be schematized with the following sequence of steps: **(1)** The descriptor set in study was centred and scaled to mean equal to zero and standard deviation equal to one. **(2)** The PCA was applied using the R package princomp and the obtained principal components (PCs) were organized in descending order of variance explained. **(3)** The PCs were used as a SVM model input following a stepwise procedure using the defined order. This method is mainly aimed at simplifying the model fitting phase, as it does still require that all variables are computed.

3.3.2.3 Model validation and evaluation

The examination of models' fitness is performed through the comparison of the experimental and predicted properties and is needed to statistically ensure that the models are sound. The proportion of variation explained by the model and the root mean squared error (RMSE) are performed to determine the goodness of fit of the model. The explained variation measures the proportion to which a model accounts for the variance of the given data set. The concept of variation explained is, in many cases, equivalent to the correlation coefficient (Spiess & Neumeyer, 2010). Nevertheless, since in QSPR/QSAR studies it is standard to use the cross-validated squared correlation coefficient (q^2), this terminology is adopted through the document. In order to validate the robustness and predictive ability of the models, all results presented are the outcome of 10-fold cross validation or out-of-bag prediction. The process of cross-validation begins with the random division of the dataset into 10-folds of compounds. One partition is removed and used as test set and the model is created from the remaining data points, this process is repeated 10 times. The validation statistics are averaged over the rounds. An external validation with an independent dataset is considered

optimal when evaluating how well the equation generalizes the data. The training set was used to derive a model that was further used to predict the properties of the test set instances, which were not used in the model development. The predictive proportion of variation explained (Q^2) by the model and the RMSE are performed to determine the external predictive ability of the model. Finally, to ensure that results were not due to chance correlations, the dependent variable for the training set was scrambled and models were built with the randomized dependent variables (Y-randomization).

3.4 Results

If you want to build a ship, don't drum up people to collect wood and don't assign them tasks and work, but rather teach them to long for the endless immensity of sea.
~ Antoine de Saint-Exupery

Several models with different settings have been tested using the techniques defined (SVM and RF): different combinations of molecular descriptors (descriptor sets A-G - see Appendix A) and dimensionality reduction/feature selection methods for different case-studies. Other approaches (e.g. Artificial Neural Networks, Clustering, Multivariate adaptive regression splines and simulated annealing) and molecular descriptors (e.g. angle strain, number of atoms shared by more than one cycle, chirality, stereoisomerism, 3D coordinates to calculate distance between atoms and compound classes) were also tried, however only the most relevant results were selected and are presented below.

To verify the importance of feature selection methods for the prediction of properties of chemical compounds the following approach was envisaged: firstly, it is necessary to assess model behaviour without any feature selection using different sets of descriptors. Secondly, different dimensionality reduction or variable selection strategies were tested in some of the case-studies, that include the use of i) SVMs with PCA for all the feature set space; ii) using GAs coupled with SVMs for feature selection; iii) use the ranked features list as produced by RFs for searching a minimal feature set to train a SVM model.

3.4.1 Case A1 - Predicting Enthalpy of formation for Hydrocarbon compounds

3.4.1.1 Model development without a feature selection or dimensionality reduction step

In order to confirm that it is possible to eliminate variables which are not informative as predictors of the property of interest, the first step is to present model results with different sets of molecular descriptors. For that purpose both SVMs and RFs were tested. Table 3.1 summarizes the results (10-fold cross validation for SVMs and out-of-the-bag cross validation for RFs) obtained for all models, comparing the performance of the models using or not a feature selection or dimensionality reduction step. The best model using the training set of 364 compounds and RFs reached a RMSE of 50.28 which corresponds to a q^2 of 0.9393 (Table 3.1). Using SVMs the best model performed with an RMSE of 44.47, corresponding to a q^2 of 0.9520 (Table 3.1). For both models the molecular descriptors in use result of a combination of descriptor sets A (Dragon), B (Openbabel) and C (CDK) (Appendix A.2). Results for the parameter set B were good considering that it uses only 8 comprehensive structural descriptors. This parameter set greatly improved the results when combined with other descriptor sets. In general the results obtained with both RFs and SVMs are of comparable quality, however, it can be denoted that RFs produce better results than SVMs for descriptor sets with a low number of descriptors while SVMs obtain better results for descriptor sets with a high number of descriptors.

3.4.1.2 Model development with a feature selection or dimensionality reduction step

Principal components analysis to reduce the dimensionality

Analysing the correlation matrix between all the variables in the dataset in study, it is possible to verify that the variable space presents a significant degree of redundancy. In order to test how the correlation between the variables affects the model performance we will use PCA to remove linear correlations and compare the results. To ensure adequate comparison of the values for each variable, each one was centred and scaled to mean equal to zero and standard deviation equal

Table 3.1: Compilation of the best results for case-study A1 using different descriptor sets and/or combinations of descriptor sets, number of variables (Nvars) after descriptor set pre-processing, the squared correlation coefficient (q^2) and the root mean square error (RMSE) for 10-fold cross-validation using Support Vector Machines or Random Forests. Different approaches to feature selection (FS) are also applied, namely: Principal Components Analysis (PCA), Genetic Algorithms (GA) and Random Forests - Variable Importance (RF-VI).

FS	**Descriptor set**					**FS technique**	**Nvars**	**ML model**	**RMSE**	q^2_{cv}
	A	**B**	**D**	**F**	**G**					
No	X	X	X				1485	RF	50.28	0.9303
	X						1273		50.40	0.9300
		X	X				212		52.72	0.9234
		X					8		55.91	0.9139
			X				204		90.97	0.7720
				X			364		92.71	0.7632
					X*		100		34.90*	0.9566*
	X	X	X				1485	SVM	44.47	0.9520
		X					8		63.81	0.8855
			X				204		78.01	0.8270
				X			364		90.47	0.7768
	X						1273		166.77	0.2443
Yes	X	X	X			**RF - VI**	**89**		**34.10**	**0.9686**
	X	X	X			PCA	28 PC		34.87	0.9671
	X	X	X			GA	58		47.10	0.9391

* Polycyclic compounds were excluded from the training set, reducing it to 236 compounds.

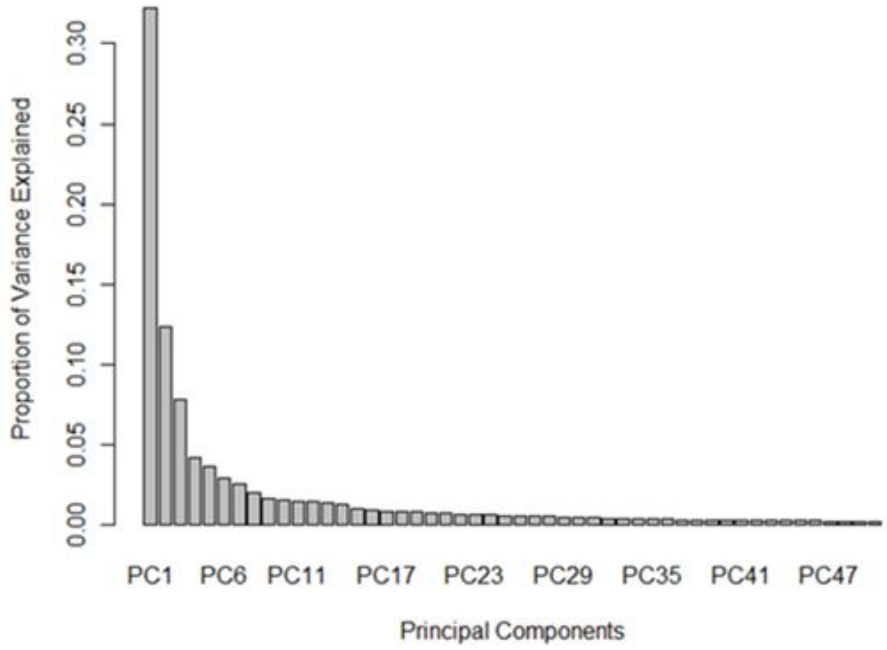


Figure 3.3: Proportion of variance in the descriptor set that is explained by each principal component (for readability the plot was truncated after the fiftieth component).

to 1.0. The plot represented in Figure 3.3 shows the proportion of variance in the dataset that is explained by each principal component (PC). The 3 first PCs are enough to explain 52.4% of the variance in the original dataset and the most significant 123 principal components are sufficient to explain 99% of the variance in the original dataset (Appendix C.1).

To use PCs as model inputs, the same question of how many components are necessary for adequate modelling is pertinent. Therefore, a stepwise approach for model construction was followed. Accordingly, several SVM models were fitted adding progressively more components following the decreasing order of the proportion of variance explained, until 150 components were present. Each model was evaluated using 10-fold cross validation. It was verified that the best model, providing the minimum RMSE (34.87), was obtained using the first 28 PCs (Table 3.1), and from this point on the prediction performance decreases for each PC added (Figure 3.4).

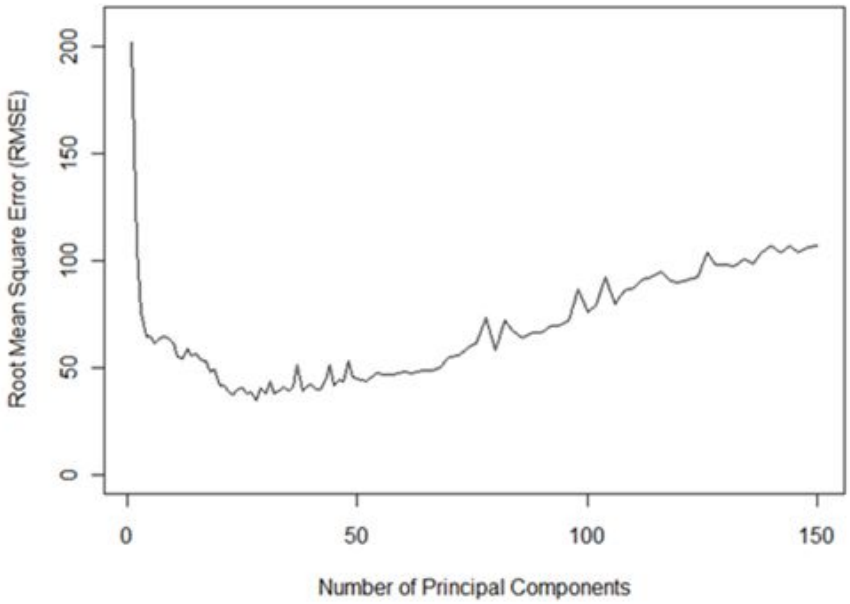


Figure 3.4: Comparison of the root mean square error (RMSE) for each predictive SVM model using an increasing number of principal components in descending order of proportion of variance explained (previously determined using principal components analysis).

Genetic algorithms for feature selection

A GA procedure for variable selection was adapted to this problem and implemented. The algorithm parameters were subjected to preliminary screening in order to ensure that the heuristic is able to adequately search variables' solution space, evaluating each set of variables found during the process with a SVM, and using the cross-validated score to rank and select each proposed subset of variables. The GA strategy that produced the best results using descriptor sets A, B and D used a population of 80 chromosomes, with a mutation rate of 2.5%, and cross over was verified as irrelevant. Initial solutions used an initial density of 4.0% meaning that, at most, 59 features are being selected for each model. During the optimization process it was verified that there were no improvements in the model performance after 1000 generations. The genetic algorithm heuristic was repeated 10 times and the final result is the average of the best solution in each run (Leardi, 2001). The obtained RMSE value was 47.10, corresponding to a

q^2 of 0.9391, using an average of 58 variables (Table 3.1). It is important to note that the list of variables selected with this method varied widely within models, with only 2 or 3 common variables per run, showing that this method although capable of producing solutions of similar quality than using all the variables, is not coherent on the set of features selected. However, it is noteworthy that approximately half of the selected descriptors are Daylight fingerprints (descriptor set D), representing certain structural fragments (Appendix C.1).

Variable importance index from Random Forests

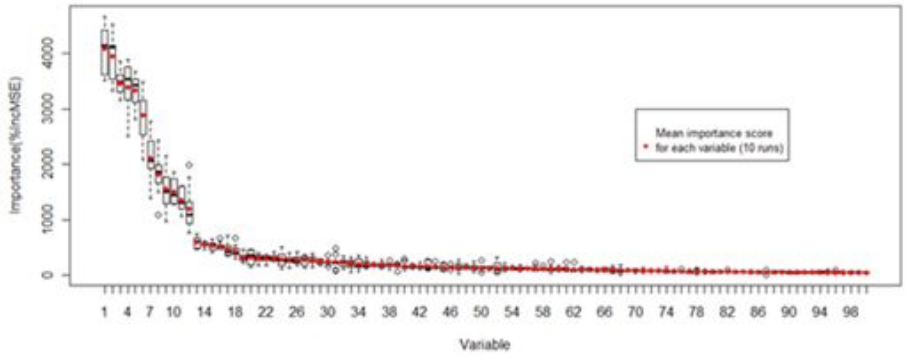


Figure 3.5: Boxplots depicting the distribution of the importance score (percentage of prediction error increased when the variable is permuted randomly) of each variable, ordered by mean importance (marked with red dots) resulting from 10 runs of the model.

In order to find the ordered list of variables according to their importance, the random forest model fitted previously was used and the importance of each variable in the final model was recorded. Due to the stochastic nature of the random forest approach, this procedure was repeated 10 times, and in the end this rank order was averaged for each variable. The variables were then sorted according to the average variable importance in descending order (Figure 3.5). These results clearly suggest that there are six very important descriptors and six moderately important ones while the others are of small importance and that the group of most important variables is not interchangeable since they have a clear difference in the quantity increased in prediction error. The results appear to be coherent and robust, with the first 20 descriptors occupying coherently the first

positions in the rank, clearly illustrating the importance of each in the current problem (Appendix C.1).

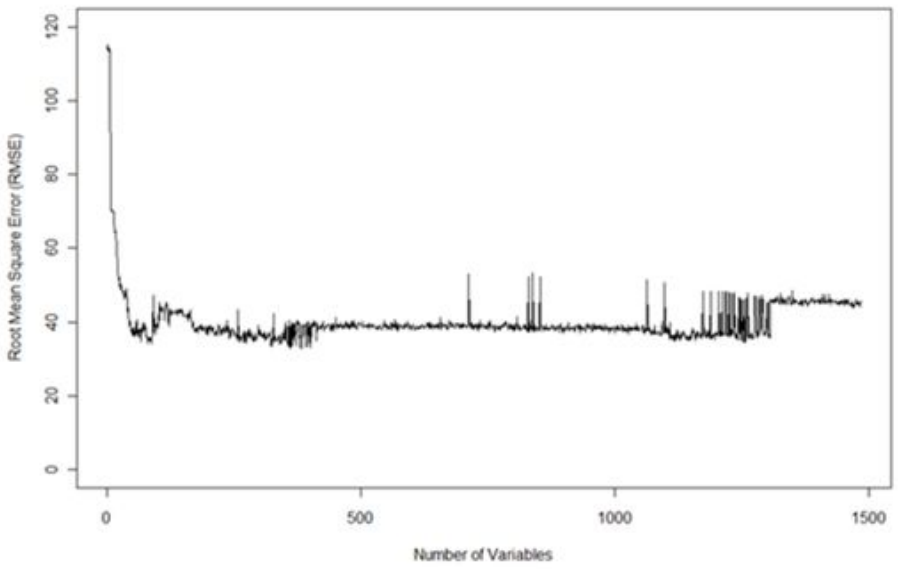


Figure 3.6: Comparison of the root mean square error (RMSE) for each predictive SVM model using an increasing number of variables in descending order of importance (previously determined using Random Forests).

With the produced descriptor rankings, the procedure followed was similar to the one used for PCA where each variable was introduced stepwise into a new model fitted with SVMs, and recording the statistical results for each new feature added. The 10-fold cross validation results for each iteration are shown in Figure 3.6 and its analysis show that a minimum RMSE (32.82) corresponding to a q^2 of 0.9706 was reached when 385 variables were used. However, it can be verified that the number of variables can be reduced to 89 without losing much predictive power, with an RMSE of 34.10 and a q^2 of 0.9686 (Table 3.1). Nonetheless, it can be verified that, in general, the predictive power of the models does not increase after 200 variables are added.

SVM models demonstrated high sensitivity to the number of input variables, and using a smaller descriptor set have, in general, better predictive power than larger descriptor sets. The use of GAs has selected descriptor sets that are able to produce good results with a limited amount of variables, yet we did not found any

coherency in the descriptors selected, which precludes the use of this technique as a reliable tool for selecting variables. PCA has produced model results that are statistically similar to the variable ranking approach as considered by RFs, yet, PCA still requires the computation of all 1485 descriptors for its application which is a relevant shortcoming. The fact that the results produced by PCA and variable ranking approach as considered by RFs are similar is an evidence, as also argued by Peterangelo & Seybold (2004), that the effects of correlation between descriptors mostly affects the interpretation of the model, with only slight effect on its predictive power. Thus the random forest based variable ranking approach is the natural choice for a final model, which, for the present problem, is able to reach robust models using only 89 molecular descriptors. The results for the scrambling runs are very poor which proves that a real relationship exists between the datasets of descriptors and the real dependent variable.

3.4.1.3 Model Validation with an Independent Validation Set

All the results presented so far have been obtained using 10-fold cross validation. It is important nevertheless to use an external and independent validation set to perform an unbiased validation of the selected model (Gramatica, 2007; Tropsha, 2010). Therefore to assess the model validity a test was performed with an independent validation set of 100 molecules, which was never considered in any of the training phases. The predictive performance of the 89-features model to this data was similar to the one obtained with 10-fold cross-validation, with an RMSE of 48.64 and a predictive proportion of variation explained (Q^2) of 0.9607. These values confirm the robustness of the approach and the effectiveness of the feature selection phase in capturing the relevant information for modelling.

3.4.2 Case A2 - Predicting Enthalpy of formation and phase change for ThermInfo's dataset

3.4.2.1 Model development without a feature selection or dimensionality reduction step

In order to confirm that it is possible to eliminate variables which are not informative as predictors of the property of interest, the first step is to present model results with all variables of different sets of molecular descriptors for all properties of this case-study - A.1.1.2. For that purpose both SVMs and RFs were tested. Table 3.2 summarizes the results (10-fold cross validation for SVMs and out-of-the-bag cross validation for RFs) obtained for all models using different descriptor sets, comparing the performance of the models using or not a feature selection step. Detailed results are available in the Appendix C.2. In general, results obtained with both RFs and SVMs are of comparable quality, however, it can be denoted that RFs produce better results than SVMs with descriptor set C due to the small number of descriptors (181) while SVMs obtain better results for descriptor sets with a high number of descriptors, namely the combination of descriptor sets C and D. The models are able to predict more accurately the Standard Enthalpy of formation in gas, liquid and crystalline phases than Standard Enthalpy of phase change. Descriptor set D (fingerprints) greatly improves the results for predicting Standard Enthalpy of formation while descriptor set C (CDK descriptors) greatly improves the results for predicting Standard Enthalpy of phase change.

3.4.2.2 Model development with a feature selection step

Variable importance index from Random Forests

In order to find the ordered list of variables according to their importance, the random forest model fitted previously was used and the importance of each variable in the final model was recorded. Due to the stochastic nature of the random forest approach, this procedure was repeated 10 times, and in the end this rank order was averaged for each variable. The variables were then sorted according to the average variable importance in descending order. With the

Table 3.2: Compilation of the best results for all properties (Standard Molar Enthalpy of Formation: crystalline (crys), liquid (liq) and gas phases and Standard Molar Enthalpy of Phase Change: fusion (phasecl), vaporization (phaselg) and sublimation (phasecg)) of case-study A2, number of compounds with experimental property (N) using different descriptor sets (C or D) and combinations of descriptor sets (C and D), number of variables (Nv) after descriptor set preprocessing, the squared correlation coefficient (q^2_{cv}) and the root mean square error (RMSE) for 10-fold cross-validation using Support Vector Machines (SVM) or Random Forests (RF). Random Forests - Variable Importance (RF-VI) for feature selection is also applied.

Prop	N	ML	Descriptor Set								
			C†			D‡			C + D§		
			Nv	RMSE	q^2_{cv}	Nv	RMSE	q^2_{cv}	Nv	RMSE	q^2_{cv}
crys	1159	RF*	181	141.9056	0.9087	987	201.6210	0.8158	1168	148.7527	0.8997
		SVM	181	202.6918	0.8358	987	188.0969	0.8405	1168	78.9200	0.9720
		RF-VI	150	72.4050	0.9765	124	179.4784	0.8541	**291**	**77.3952**	**0.9731**
liq	1186	RF*	181	71.0040	0.9730	987	137.8466	0.8984	1168	73.2213	0.9713
		SVM	181	190.7609	0.8747	987	126.7258	0.9165	1168	72.2058	0.9728
		RF-VI	127	44.2841	0.9895	184	123.2143	0.9195	**236**	**43.3661**	**0.9899**
gas	1391	RF*	181	76.2936	0.9653	987	137.0411	0.8880	1168	80.5329	0.9613
		SVM	181	179.8249	0.8580	987	123.5806	0.9097	1168	46.4523	0.9871
		RF-VI	106	54.7530	0.9822	232	117.9896	0.9172	**265**	**46.0746**	**0.9874**
phasecl	63	RF*	181	5.3601	0.8481	987	8.3195	0.6342	1168	5.4990	0.8401
		SVM	181	5.7935	0.8271	987	9.3256	0.6342	1168	5.2646	0.8542
		RF-VI	16	4.8777	0.8744	50	8.3037	0.6399	**29**	**4.8358**	**0.8766**
phaselg	893	RF*	181	5.2990	0.8975	987	9.3196	0.6830	1168	5.3070	0.8972
		SVM	181	4.7548	0.9175	987	9.2768	0.6890	1168	4.5829	0.9234
		RF-VI	110	4.5621	0.9242	247	9.0805	0.7021	**227**	**4.4806**	**0.9268**
phasecg	464	RF*	181	33.6805	0.4194	987	36.3050	0.3253	1168	34.0101	0.4079
		SVM	181	36.8117	0.3087	987	37.3530	0.2874	1168	24.3097	0.6987
		RF-VI	112	25.3629	0.6799	50	31.5467	0.4959	**225**	**23.6171**	**0.7149**

* RFs of 500 trees. The mtry values are specified in Appendix C.2.
† Descriptor set C (CDK descriptors) has 196 descriptors of which 181 have variance different from zero.
‡ Descriptor set D (Fingerprints) has 1023 descriptors of which 987 have variance different from zero.
§ Descriptor sets C and D (CDK descriptors and Fingerprints) have 1219 descriptors of which 1168 have variance different from zero.

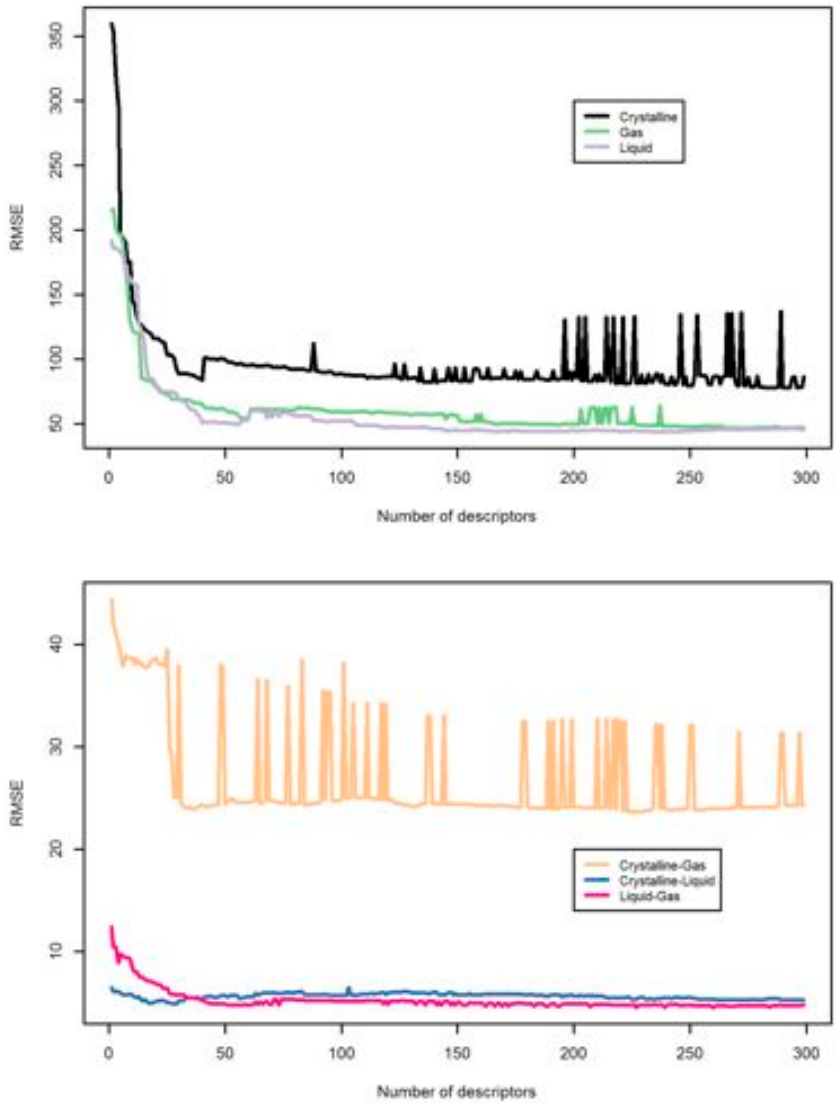


Figure 3.7: Comparison of the root mean square error (RMSE) for each predictive SVM model using an increasing number of variables (descriptor sets C and D) in descending order of importance (previously determined using Random Forests) for all properties in case-study A2 (top plot presents results for Standard Enthalpy of formation while lower plot presents results for Standard Enthalpy of phase change).

produced descriptor rankings, each variable was introduced stepwise into a new model fitted with SVMs, and recording the statistical results for each new feature

added. The 10-fold cross validation results for each iteration are shown in Figure 3.7 using the combination of descriptor sets C and D and its analysis show that a minimum RMSE is obtained around 200 variables except for standard enthalpy of phase change crystalline-liquid phases. However, it can be verified that the number of variables can be reduced to approximately 40 to 50 variables without losing much predictive power.

SVM models showed to be sensitive to the number of input variables, and using a smaller descriptor set have, for all properties, better predictive power than larger descriptor sets. Thus the random forest based variable ranking approach is the natural choice for a final model, which, for the present problem, is able to reach robust models using a smaller number of molecular descriptors. The results for the scrambling runs are very poor which proves that a real relationship exists between the datasets of descriptors and the real dependent variable.

3.4.2.3 Model Validation with an Independent Validation Set

Table 3.3: Compilation of the best results obtained using the best model for each property (Standard Molar Enthalpy of Formation: crystalline (crys), liquid (liq) and gas phases and Standard Molar Enthalpy of Phase Change: fusion (phasecl), vaporization (phaselg) and sublimation (phasecg)) in an independent test set of case-study A2, number of compounds in the test set (N) using a combination of descriptor sets C and D, number of variables (Nvars) determined in the modelling phase, the squared correlation coefficient (Q^2) and the root mean square error (RMSE) using Random Forests - Variable Importance (RF-VI) for feature selection.

Property	**N**	**Nvars**	**ML**	**RMSE**	Q^2
crys	300	291	RF-VI*	60.7064	0.9780
liq	300	236	RF-VI*	26.0536	0.9970
gas	350	265	RF-VI*	27.8704	0.9936
phasecl	20	29	RF-VI*	3.2085	0.9093
phaselg	200	227	RF-VI*	2.4233	0.9790
phasecg	150	225	RF-VI*	6.3806	0.9351

* RF-VI using a combination of descriptor sets C and D.

All the results presented so far have been obtained using 10-fold cross validation. It is important nevertheless to use an external and independent validation set to perform an unbiased validation of the selected model (Gramatica, 2007; Tropsha, 2010). Therefore to assess the model validity, it was tested with test sets, which were never considered in any of the training phases. The predictive performance of the selected best models in the training phase in the test set are presented in Table 3.3. These results show that the predictive performance in the test set is similar or better than the obtained with 10-fold cross-validation. These values confirm the robustness of the approach and the effectiveness of the feature selection phase in capturing the relevant information for modelling.

3.4.3 Case D - NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge

3.4.3.1 Model development without a feature selection or dimensionality reduction step

This case-study is related with the participation in the 8th Dialogue on Reverse Engineering Assessment and Methods (DREAM 8) Challenges (June – September, 2013) which is thoroughly described in Appendix A.1.4. The goal of this NIEHS-NCATS-UNC DREAM toxicogenetics challenge is to model population-level cytotoxicity parameters to unknown chemical compounds based on chemical structure attributes. For that propose, it is necessary to model the median, 5th quantile, and 95th quantile EC_{10} for each of 106 compounds in the training cytotoxicity data file as estimated from all 884 cell lines using chemical descriptors. In a similar way to that carried out in the previous case studies, we will confirm that it is possible to eliminate variables which are not informative as predictors of the property of interest to improve model robustness and predictive performance. For that purpose both SVMs and RFs were tested with and without feature selection. Table 3.4 summarizes the results (10-fold cross validation for SVMs and out-of-the-bag cross validation for RFs) obtained for all the models, comparing the performance of the models using or not a feature selection or dimensionality reduction step. This table presents the models to predict the median EC_{10}, however the remaining parameters show similar results. The best model using the

training set of 106 compounds and RFs reached a RMSE of 0.5658 which corresponds to a q^2 of 0.0623 (Table 3.4). Using SVMs the best model performed with an RMSE of 0.5464, corresponding to a q^2 of 0.1293 (Table 3.4). Both models are not able to predict the toxicology of chemical compounds.

Table 3.4: Compilation of the best results for case-study D using different descriptor sets and/or combinations of descriptor sets, number of variables (Nvars) after descriptor set pre-processing, the squared correlation coefficient (q^2_{cv}) and the root mean square error (RMSE) for 10-fold cross-validation using Support Vector Machines (SVM) or Random Forests (RF). Different approaches to feature selection (FS) are also applied, namely: Principal Components Analysis (PCA) and Random Forests - Variable Importance (RF-VI).

FS	Descriptor set				FS technique	Nvars	ML model	RMSE	q^2_{cv}
	A	C	D	E					
No			X			752	RF	0.5658	0.0623
				X		1580		0.6022	0.0000
			X	X		2332		0.5875	0.0000
		X				157		0.5887	0.0000
	X					1666		0.5993	0.0000
		X	X	X		2489		0.6000	0.0000
			X			752	SVM	0.5464	0.1293
	X					1666		0.5773	0.0490
	X	X	X			2575		0.5758	0.0331
		X				157		0.5840	0.0016
Yes			X		**RF-VI**	**95**		**0.4133**	**0.5046**
				X	RF-VI	109		0.4676	0.3681
	X	X	X		RF-VI	41		0.5281	0.1942
				X	PCA	8		0.5333	0.1840
	X	X	X	X	PCA	4		0.5409	0.1437

3.4.3.2 Model development with a feature selection or dimensionality reduction step

Principal components analysis to reduce the dimensionality

The plot represented in Figure 3.8 shows the proportion of variance in the dataset D that is explained by each PC using all descriptor sets (A, C, D and E). The 4 first PCs are enough to explain only about 10% of the variance in the original dataset, 40 PCs explain about 50% of the variance and 104 PCs are needed to explain about 99% of the variance. Once again, a stepwise approach for model construction was followed, until 105 components were present. Each

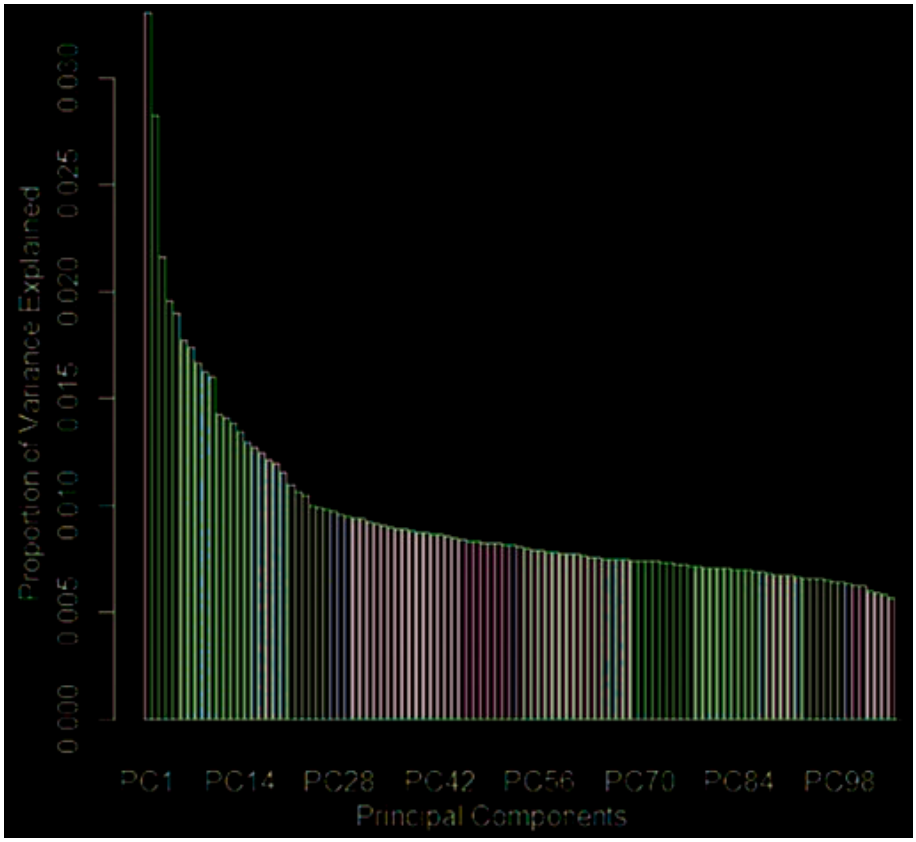


Figure 3.8: Proportion of variance in the descriptor set (A + C + D + E) that is explained by each principal component (for readability the plot was truncated after the one-hundredth component).

model was evaluated using 10-fold cross validation. It was verified that the best model, providing the minimum RMSE (0.5409), was obtained using the first 4 PCs (Table 3.1), and from this point on the prediction performance decreases for each PC added (Figure 3.9).

Variable importance index from Random Forests

In order to find the ordered list of variables according to their importance, the random forest model fitted previously was used and the importance of each variable in the final model was recorded. The importance of each variable is in general equally low.

With the produced descriptor rankings, the procedure followed was similar to the one used for PCA where each variable was introduced stepwise into a new

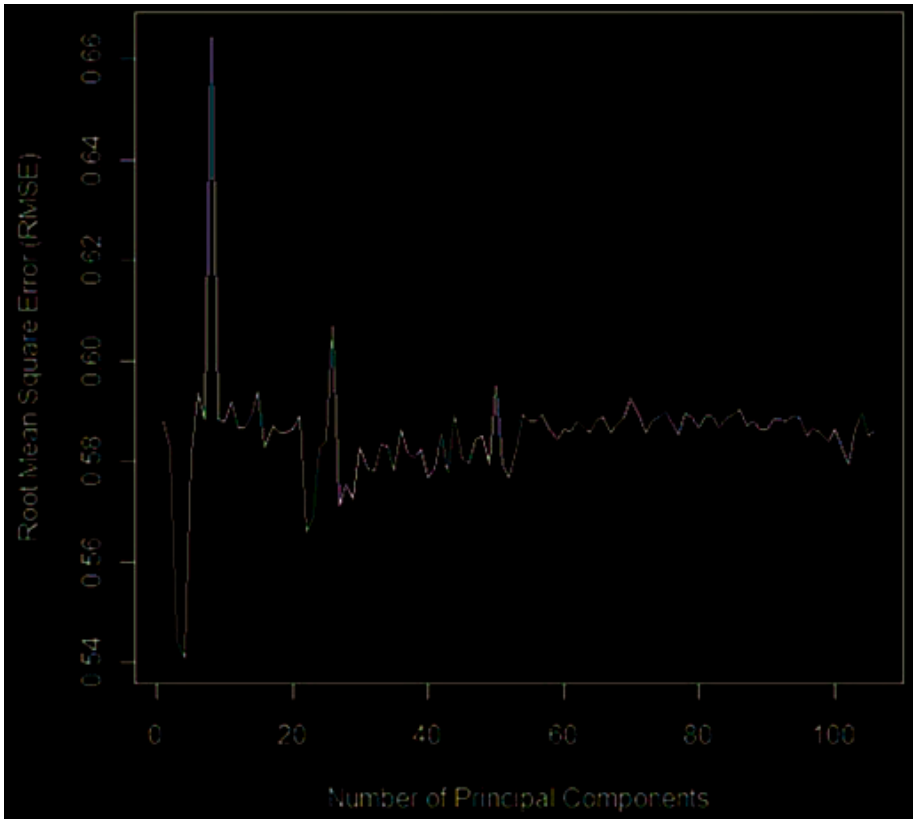


Figure 3.9: Comparison of the RMSE for each predictive SVM model using an increasing number of principal components in descending order of proportion of variance explained (previously determined using principal components analysis).

model fitted with SVMs, and recording the statistical results for each new feature added. The 10-fold cross validation results for each iteration are shown in Figure 3.10 and its analysis show that a minimum RMSE (0.4133) corresponding to a q^2 of 0.5046 was reached when 95 variables were used.

SVM models showed to be sensitive to the number of input variables, and using a smaller descriptor set have clearly a better predictive power than larger descriptor sets even when the dimensionality of the descriptors is reduced using PCA. Thus the random forest based variable ranking approach is the natural choice for a final model, which, for the present problem, is able to reach predictive models using only 95 molecular descriptors. The results for the scrambling runs are very poor which proves that a real relationship exists between the datasets of descriptors and the real dependent variable.

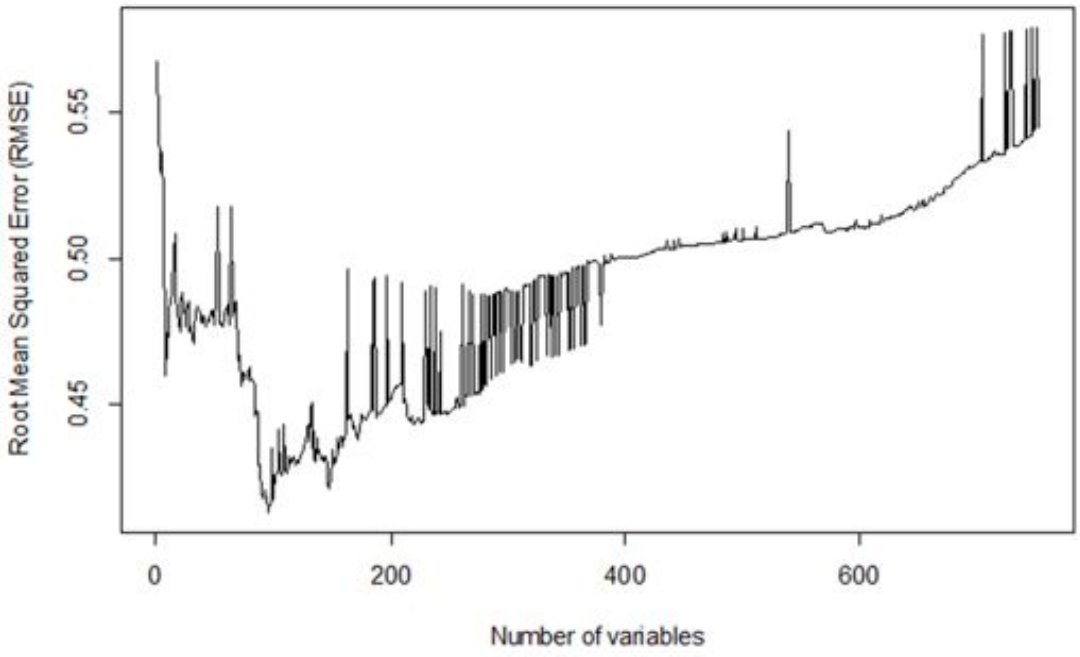


Figure 3.10: Comparison of the root mean square error (RMSE) for each predictive SVM model using an increasing number of variables in descending order of importance (previously determined using RFs) for case-study D.

The details about the models that were submitted to this challenge are presented in the following link: https://www.synapse.org/#!Synapse:syn2219104/wiki/.

3.4.3.3 Model Validation with an Independent Validation Set

To assess the best model validity and evaluate model submissions to this challenge, an independent validation set of 50 molecules was used, which was never considered or known in any of the training phases. The predictive performance of the 95-features model to this data was an RMSE of 1.0977 and a predictive proportion of variation explained (Q^2) of 0.4050.

3.4.4 Case G - Blood-Brain Barrier (BBB) Penetration Modeling

In silico modelling of BBB permeation is extremely important for designing molecules targeting the central nervous system as well as for avoiding possible side effects. Several studies in the literature have attempted to predict BBB penetration, so far with limited success and few, if any, application to real world drug discovery and development programs, since only about 2% of small molecules can cross the BBB and smaller molecules (smaller molecular weight (MW) have higher probability to cross the barrier, however available data sets over-represent the molecules that show an ability to permeate the BBB. Following a previously proposed Bayesian model on BBB penetration developed by our research team (Martins *et al.*, 2012) (`http://b3pp.lasige.di.fc.ul.pt`) to overcome this problem by differentially sampling the available data using a Bayesian approach in order to approximate what is expected for any molecule for which there is no previous knowledge, the aim of this new study is to ascertain the minimum subset of descriptors needed to predict the BBB penetration with a good performance, less computational/time cost and in a more robust way, since the presence of irrelevant or redundant features can cause poor generalization capacity.

The dataset for model training and cross-validation was comprised of 1850 molecules (Appendix B.6) for which molecular descriptors were calculated and pre-processed and it is thoroughly described in Appendix A.1.7. An unbiased dataset should approximate the population statistics and could be produced by randomly sampling from a pool of molecules of unknown BBB penetration. The current dataset with a stronger component of BBB_+ molecules is clearly biased. The importance of an unbiased sample cannot be underestimated in prediction modelling. If the training dataset does not represent the population, the fitted model also will not be adequate and will be biased in the same direction of the training set.

To solve this problem, an approach proposed by Martins *et al.* (2012) was followed that aimed to produce an unbiased training dataset from the currently biased dataset. The basic approach followed for model fitting used a differential

sampling according to the a priori probabilities of each molecule to belong to BBB_+ or BBB_-:

1. From the full dataset S select randomly a training set T of arbitrary size.

2. Use the molecules of training set T to compute the prior probability density function (PDF) according to molecular weight ($P(m_{MW}|h_+)$ and $P(m_{MW}|h_-)$ where $m \in T$

3. Using the molecules in T, proceed iteratively until a user-specified number of instances (N) for a re-sampled new training set (T') is selected:

   (a) From T select randomly with reposition one instance and check its observed class θ: $\theta = (+)$ or $\theta = (-)$ for BBB_+ or BBB_- respectively

   (b) According to the MW of the compound get its appropriate likelihood ($P(m_{MW}|h_\theta)$ from the generated prior PDF

   (c) Generate a random number r and calculate $P(h_\theta|P(m_{MW})$ according to equations (3.2) and (3.3). Thus, the a priori $P(h_+)$ probability is in fact the probability that a compound with a given MW can cross the BBB ($P(h_+|m_{MW})$). This can be calculated by using Bayes theorem with the probability distribution function of all compounds that penetrate the BBB ($P(m_{MW}|h_+)$), the *a priori* knowledge of how the probability that an unknown compound may cross the BBB ($P(h_+)$) and the probability of occurrence of a given MW ($P(m_{MW})$). These latter values are identical for BBB_- and BBB_+ molecules and thus can be discarded.

$$P(h_+) = P(h_+|m_{MW}) = \frac{P(m_{MW}|h_+)P(h_+)}{P(m_{MW})} \tag{3.2}$$

$$P(h_-) = P(h_-|m_{MW}) = \frac{P(m_{MW}|h_-)P(h_-)}{P(m_{MW})} \tag{3.3}$$

   (d) If $r \leq P(h_\theta|m_{MW})$ update set T' adding the instance selected in (a)

This procedure will produce a training set that will mirror more adequately the real world molecule space, as each molecule will be selected according to its a priori probability. The above method is further able to produce arbitrarily large training sets as each molecule may be selected more than once because the sampling process is constructed with reposition. One inevitable consequence of this differential selection procedure is that if the size (N) of the re-sampled training set (T') is not large enough, some molecules may never be selected for model building. The process followed used the hybrid two-phase approach to select descriptors described in this chapter. Because of the sampling process, some changes were performed in the methodology: after variable importance calculation using all instances, different number of variables n (with $n = 20, 50, 100, 200, 300, 500, 800, 1057$ for descriptor sets B and D and $n = 20, 50, 100, 200, 500, 1500, 2569$ for descriptor sets A, B and D) will be used for model building and for each number of variables the process of sampling and model building will be repeated 300 times. Figure 3.11 shows the workflow of the methodology adaptation for this classification problem.

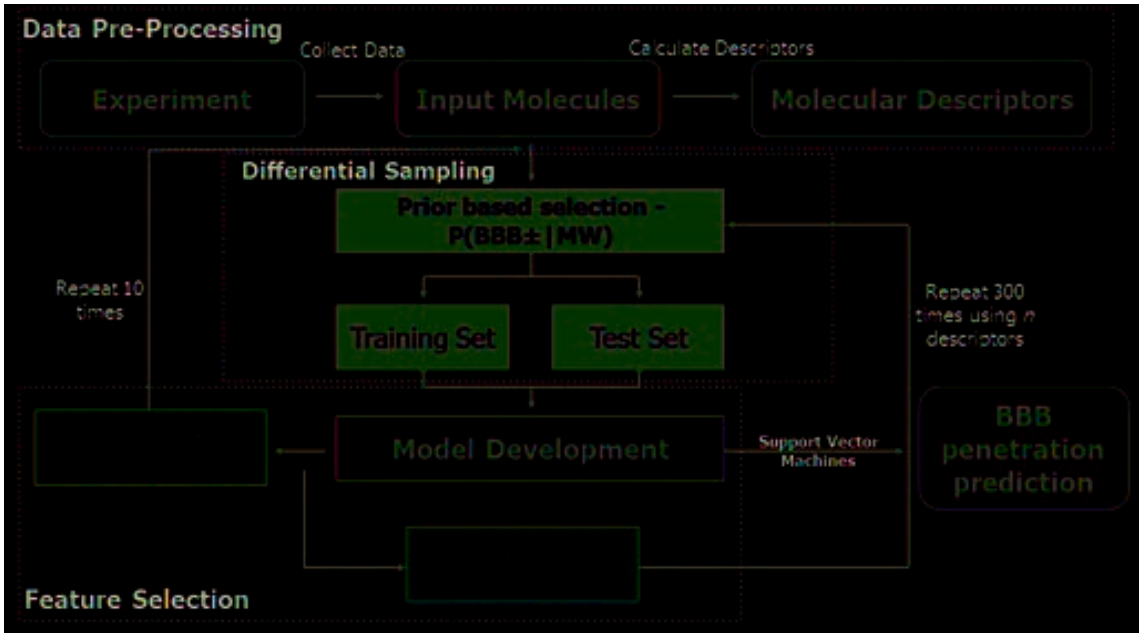


Figure 3.11: General workflow of the hybrid approach using Random Forests for Feature Selection for predicting properties of compounds based on molecular descriptors adapted to BBB permeation prediction.

Figure 3.12 present the variables sorted according to the average variable importance in descending order. For descriptor sets B + D there are clearly three important variables that distinguish from all others. For descriptor sets A + B + D, there are 7 very important variables, however, contrarily to the most important variables in the previous combination of descriptors, these variables are not as distinguishable from the rest. Details about these lists of variables are available in Appendix C.3. The results appear to be coherent and robust, with the first 20 descriptors occupying coherently the first positions in the rank, clearly illustrating the importance of each in the current problem.

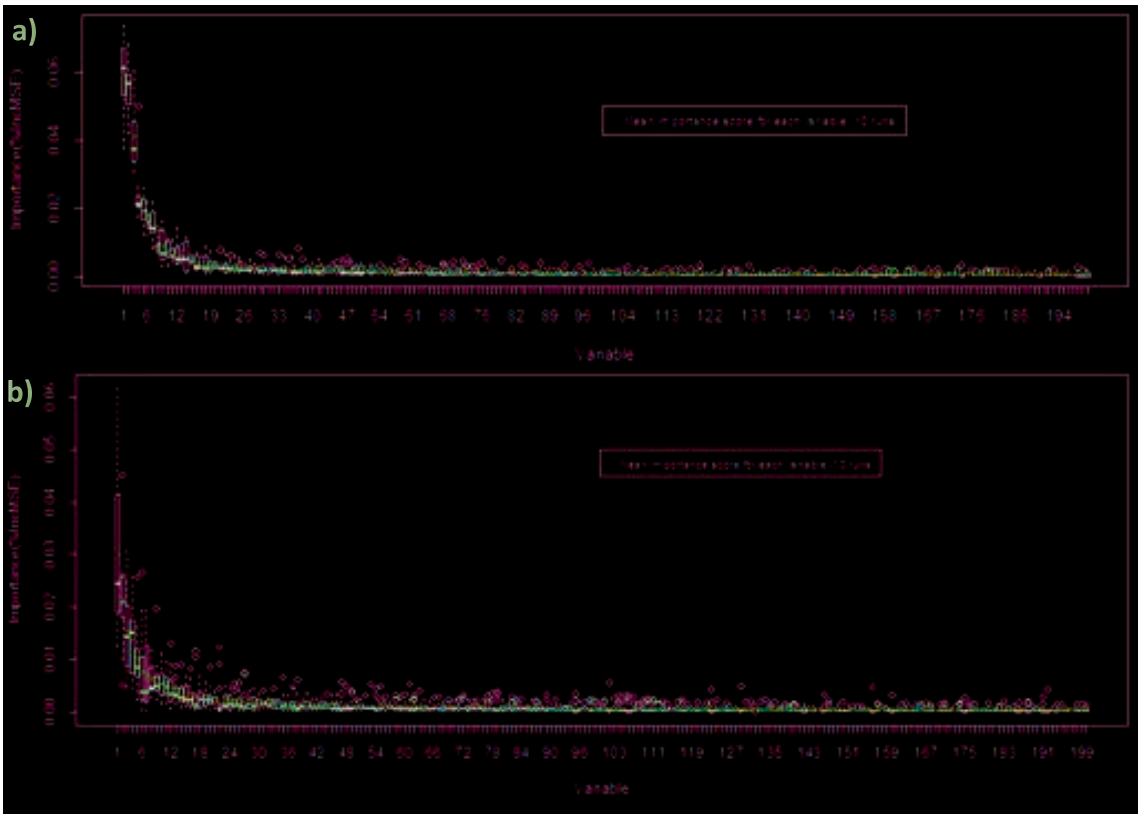


Figure 3.12: Boxplots depicting the distribution of the importance score (percentage of prediction error increased when the variable is permuted randomly) of each variable, ordered by mean importance (marked with red dots) resulting from 10 runs of the model: **a)** Descriptor sets B + D and **b)** Descriptor sets A + B + D. For readability these plots were truncated after the 200^{th} variable.

Table 3.5 summarizes the performance of the models in terms of accuracy and

Matthews correlation coefficient (MCC or ϕ) using different number of variables for both combination of descriptor sets. More details can be found in Appendix C.3.

Table 3.5: Compilation of the best results for case-study G using different combinations of descriptor sets (see Appendix A.2), number of variables (Nvars) after descriptor set pre-processing, expected accuracy (Acc) and the expected Matthews correlation coefficient (ϕ) for 300 model repetition using Random Forests - Variable Importance.

Descriptor Sets	**B + D**		**A + B + D**	
Nvars	**Expected ϕ**	**Expected Acc (%)**	**Expected ϕ**	**Expected Acc (%)**
20	0.489	91.5	**0.591**	**93.5**
50	0.540	90.3	0.535	91.4
100	0.648	91.5	0.507	88.3
200	**0.687**	**90.1**	0.513	88.7
300	0.482	86.9	-	-
500	0.457	86.0	0.539	88.8
800	0.424	82.0	-	-
1057	0.431	88.6	-	-
1500	-	-	0.508	87.4
2569	-	-	$8.680E^{-06}$	10.0

Results presented in Table 3.5 strongly suggest that using a selected smaller number of chemical descriptors is better than using all available information and produces significantly better models. Nonetheless simplistic models with too few descriptors are not enough to produce reliable results. The optimal number of descriptors was found to be around 200 using descriptor sets B and D, which produced cross-validated results with an expected Mean Squared Contingency Coefficient (MSCC) of 0.687, and overall accuracy of 90.1%, thus definitely superior to the results obtained when using all descriptors: MSCC=0.431 and accuracy = 88.6%.

Figure 3.13 present density plots depicting the distribution of the ϕ and overall accuracy with different number of tested variables (ranked according to variable importance) using descriptor sets B + D. For readability the plots presenting the results for the best model (with number of variables = 200) are also presented separately. These plots corroborate previous observations that models with a smaller number of descriptors are able to obtain better models.

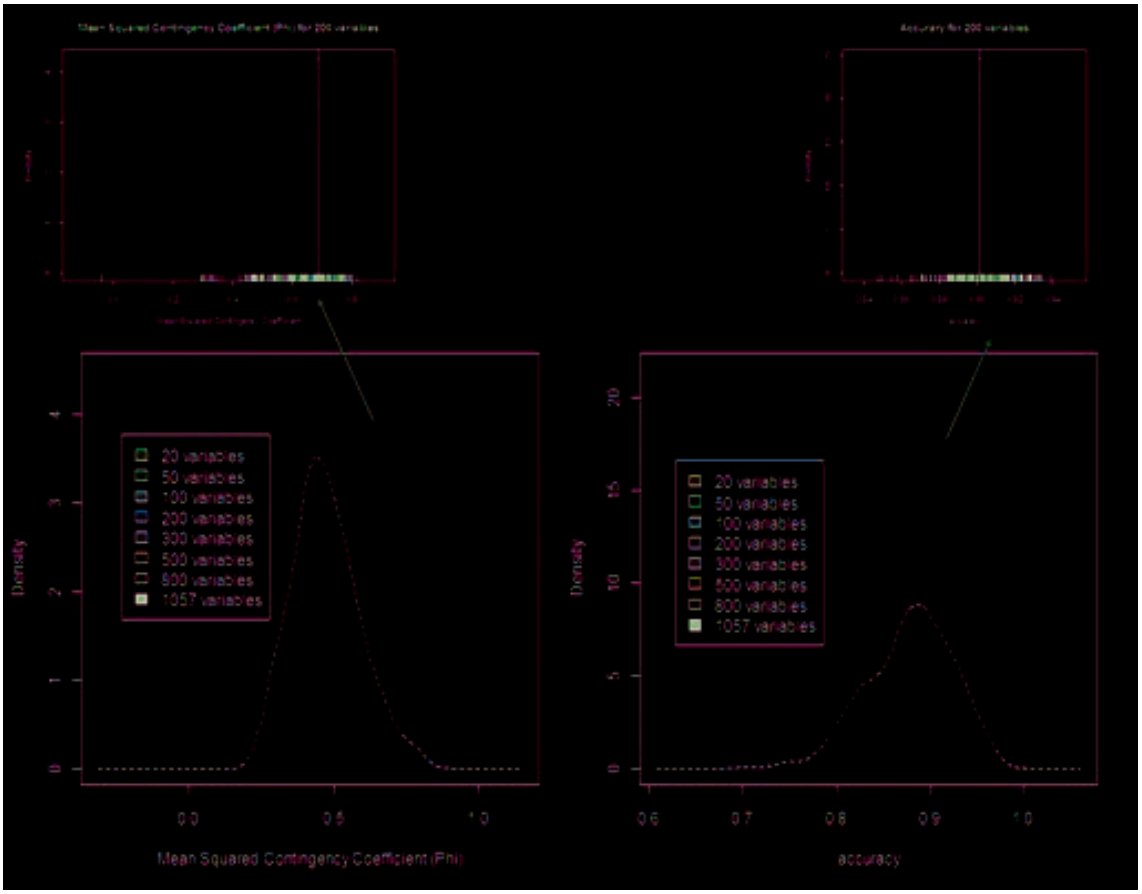


Figure 3.13: Density plots depicting the distribution of the ϕ and overall accuracy with different number of variables (ranked according to variable importance) using descriptor sets B + D. For readability the plots presenting the results for the best model (number of variables = 200) are also presented separately.

3.4.5 Web-based Information Systems and Tools

3.4.5.1 ThermInfo: Collecting, Retrieving, and Estimating Reliable Thermochemical Data

Standard enthalpies of formation are used for assessing the efficiency and safety of chemical processes in the chemical industry. However, the number of compounds for which the enthalpies of formation are available is many orders of magnitude smaller than the number of known compounds. Thermochemical data prediction methods are therefore clearly needed. Several commercial and free chemical databases are currently available, the NIST WebBook being the most used free

source. To overcome this problem a cheminformatics system was designed and built with two main objectives in mind: collecting and retrieving critically evaluated thermochemical values, and estimating new data based on the results of this work. In its present version, ThermInfo allows the retrieval of the value of a thermochemical property (Figure 3.14), such as a gas-phase standard enthalpy of formation, by inputting, for example, the molecular structure or the name of a compound. The same inputs can also be used to estimate data for compounds that are not in the database.

Figure 3.14: Composite screenshot example of some data retrieval features in the ThermInfo Web information system. **(1)** Example of three types of data input: **(a)** Quick Search, term-based search; **(b)** Advanced Search, multiple search fields based on specific structural characteristics; **(c)** Structural or Substructure Search, based on the molecular structure drawn in a Java applet. **(2)** Search result list: the query description and the list of compounds found in the database. **(3)** Detailed information available for the selected compound.

The system development started during my master's thesis work (ThermInfo 1.0) and several interdependent and complementary phases of the development of an information system were carried out, such as: (1) problem and requirements

analysis including interviews with potential users and analysis of the problems handling chemical data; (2) design the functionality of the system using low-fidelity prototypes and Unified Modeling Language (UML) class diagrams; (3) system implementation with a suitable architecture to store, manage, retrieve and predict structural and thermochemical properties of chemical compounds on the web; (4) system evaluation in terms of usability of the interface and potential problems with and without users; and (5) system maintenance with continuous monitoring of the system performance and development of enhancements and modifications. In the last few years the focus was the maintenance of the system and the implementation of new cheminformatics features. To maintain the system, ThermInfo was moved and installed in a dedicated server in two different virtual machines: one for testing and another one to serve on the web, all the code files were cleaned up, some software bugs were corrected and documentation was added. Some new features to handle the particularities of chemical data were implemented, namely similarity search using the compound SMILES or SMARTS, search by structure or substructure of a molecule using an applet to draw the structure, searching a compound by name will not only query for the main name like before but also for synonyms and names that sound like the query name, search by property range, a molecular weight calculator by name or SMILES and property prediction using ELBA's method by SMILES, compound name or structure. ThermInfo 1.9 was released for public access on the Web in March, 2011. ThermInfo will also be used to implement and make publicly available new estimation methods based on data-mining techniques developed during this work. From March, 2011 to June, 2014 this system has had over 2600 accesses from over 800 users from 57 different countries.

In terms of technologies options for the system implementation, the database is implemented using MySQL. Many of the basic application tools, scripts, and web interfaces, were developed using Hypertext Preprocessor (PHP), a server-side programming language designed especially for the web, with the possibility to be embedded in Hypertext Markup Language (HTML) code. PHP is very well documented, supports intensive transactions, runs fast, and works well with other programming languages chosen for the development of this project, MySQL and JavaScript. Both data presentation and property prediction features allow

the user to draw the chemical structure in a JAVA applet (JChemPaint) and export it as a SMILES or 3D MDL MOL file. The use of JavaScript allows a dynamic interaction with the structure. A drawn chemical structure can be converted to a downloadable file format using a Python library (OASA). Conversions between a given structure identifier and another structure identifier or representation are made using a structure-name lookup, the Chemical Identifier Resolver (NCI/CADD CIR, 2011) provided by the NCI/CADD group, using a simple Uniform Resource Locator (URL) Application Programming Interface (API) scheme or Open Babel (O'Boyle *et al.*, 2011). Pybel (O'Boyle *et al.*, 2008) (a Python library that provides access to the Open Babel toolkit) was also used to convert file formats, calculate molecular fingerprint to compare molecules, and to access data and information about structural attributes of the molecule in order to extract the ELBA parameters (see Appendix A.2.6) and predict thermochemical properties. The web interface is delivered using the open-source Apache web server. Control over access to administrative functions is performed by using Apache Hypertext Access.

The information system is publicly available at `http://therminfo.lasige.di.fc.ul.pt`. ThermInfo's strength lies in the data quality, availability (free access), search capabilities, and, in particular, prediction ability, based on a user-friendly interface that accepts inputs in several formats.

3.4.5.2 B3Info – An information system for molecular Blood-Brain Barrier penetration data

B3Info implements an information system for storage, search and retrieval of chemical molecules as related to their permeation properties to the Blood Brain Barrier. The initial molecular information that populated the database came from the published literature and was curated manually. Two types of information are present: either simple information mentioning whether a molecule is known to cross the barrier, as well as quantitative permeation information, expressed as the LogBB, is provided when available. The database stores not only the chemical data but also the original references from which the data was collected, as well as links to other chemical repositories on the web, namely: DrugBank, NCBI's PubChem, CheBI or ChemSpider for each molecule. The system further links to

the B3PP tool which implements a model for inferring blood brain barrier penetration for any molecule (Martins *et al.*, 2012). Presently this database contains data for over 1900 molecules. B3Info provides full web access to the database, allowing extensive searching capabilities. These include searching molecules according to their common name or their structural representation in the form of SMILES or InChI identifiers. It is further possible to search molecules using structural similarity, thus retrieving molecules that share similar structures, and also allowing for strict sub-structural searches. B3Info further provides an interface for drawing a graphical representation of any molecule and search for similar structures. This information system used as base the architecture of ThermInfo and was implemented by a master student, Luís Pinheiro. B3info is open and free to use and the full molecular database can be downloaded in the following link: http://b3info.lasige.di.fc.ul.pt.

3.5 Discussion

All models are wrong, but some are useful.
~ George Box

3.5.1 Case A1 - Predicting Enthalpy of formation for Hydrocarbon compounds

3.5.1.1 Selected chemical descriptors

Different feature selection or dimensionality reduction techniques were applied to select the most important descriptors in order to predict the property of interest. The stability of these methods is very important, since ideally, in the same conditions, different runs of each method should not influence the feature subset selection. The most important descriptors selected by the three methods are very different between each other, however the descriptor average molecular weight (AMW) appears as important to both GAs and variable importance calculated by RFs. GAs select mostly Daylight fingerprints, while variable importance calculated by RFs give more importance to the 2D and 3D descriptors calculated by E-DRAGON. In terms of stability, GAs are not coherent on the set of features

selected since, in general, only 2 or 3 variables are common per run while using variable importance calculated by RFs the list of most important descriptors is coherent. It is difficult to assess the relative importance/contribution of each variable in the principal components calculated by principal components analysis. Appendix C.1 makes available the lists of descriptors selected using these different selection/reduction methods.

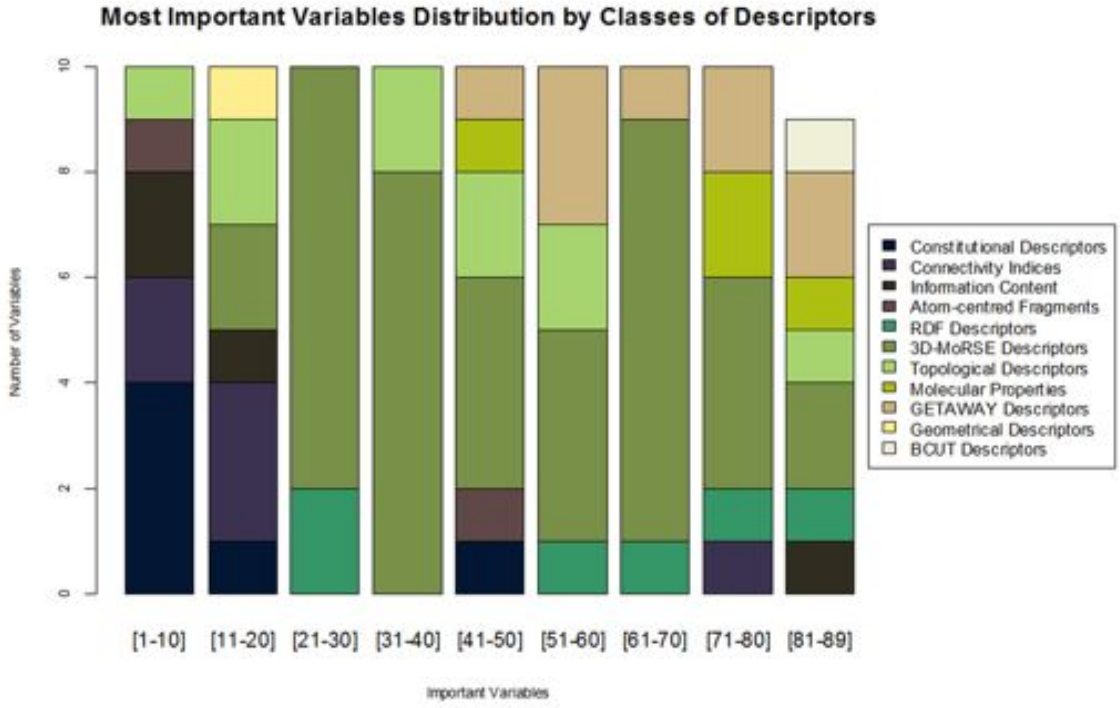


Figure 3.15: Distribution of the 89 most important variables by classes of descriptors.

The 89 most important descriptors selected using variable importance calculated by RFs were individually analysed. In a first step these were grouped into general classes (Figure 3.15). These descriptors are derived from different models and approaches, but they can be loosely grouped according to their information content: a) Constitutional descriptors, reflecting the molecular constitution and independent from molecular connectivity and conformations; b) Connectivity indices and Topological descriptors, reflecting the topology of a given structure, calculated from the vertex of the atoms in the H-depleted molec-

ular graph; c) Information content indices, reflecting the neighbourhood of an atom and edge multiplicity; d) BCUT descriptors, reflecting atomic properties relevant to intermolecular interactions, calculated from the eigenvalues of the adjacency matrix; e) Atom-centred fragments, reflecting the presence of a set of defined structural fragments; f) Radial Distribution Function (RDF) descriptors, reflecting the molecular conformation/geometry in 3D, based on the distance distribution in the molecule; g) 3D-Molecule Representation of Structures based on Electron diffraction (MoRSE) descriptors, reflecting 3D information based on the 3D coordinates of the atoms by using the same transformation as in electron diffraction; h) GEometry, Topology and Atom-Weights AssemblY (GETAWAY) descriptors, reflecting the 3D molecular geometry provided by the leverage matrix of the atomic coordinates; i) Geometrical descriptors, reflecting the conformation of a molecule based on their geometry; j) Molecular Properties, calculated using models or semi empirical descriptors (Todeschini *et al.*, 2008).

Although the 10 most important variables reflect mainly 2D information (constitutional, connectivity, information content and atom-centred fragments descriptors), the most common type of descriptors, with 40 variables, reflects 3D information (3D-MoRSE descriptors). The most important variable found for the prediction of the standard enthalpy of formation in gas phase is the average molecular weight, which represents the sum of the atomic weights of the atoms in the molecule divided by the number of atoms in the molecule (including hydrogen atoms). Unlike the molecular weight, this descriptor does not give an idea of the size of the molecule, but about the branching, type of atoms and bonds and therefore it has a good capacity to distinguish different families of hydrocarbons. Contrasting to the sets of variables selected by the model trained with GAs, which have a high accounting for fingerprints, this set of variables does not contain fingerprints.

3.5.1.2 Prediction errors analysis

The experimental values of enthalpy of formation in gas phase (kJ/mol) were compared to the predicted values using the independent validation set and represented in a scatter plot, with an RMSE of 48.64 and a Q^2 of 0.9607 (Figure 3.16 -

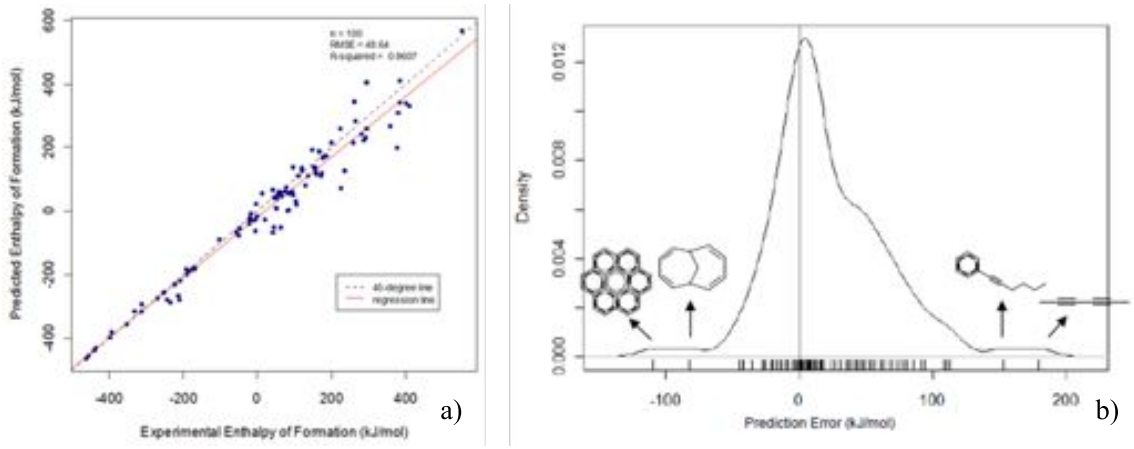


Figure 3.16: **a)** Plot of experimental versus predicted values of enthalpy of formation in gas phase (kJ/mol) using the independent validation set. **b)** Density plot of the differences between the observed values and the predicted values using the independent validation set. The structure of the compounds with most extreme prediction errors are indicated, the positive errors correspond to compounds with triple bonds (hexa-2,4-diyne and hex-1-ynylbenzene) and the negative errors correspond to compounds with more than one cycle (coronene and bicyclo[4.4.1]undeca-2,4,7,9-tetraene).

a). The majority of the data points are concentrated around the line of equality between the experimental and predicted value of the property (45-degree line) therefore, the relationship between them is strong. The distance of each symbol from the 45-degree line corresponds to its deviation from the related experimental value. The regression line indicates that generally the model predicts values close to the equality with a small deviation showing that the model is predicting with smaller values than the observed ones. The prediction errors obtained for the independent validation set were then further analysed and are represented in the Figure 3.16b). Similarly to what has already been observed, the model is predicting the enthalpy of formation with a left bias (smaller values than expected) and the most probable error is 4.10. The compounds with higher errors are the alkynes, probably due to the fact that this type of compounds are over-represented in the validation set with 12 compounds while only 4 alkynes exist in the training set and the latter is more than 3.5 times larger than the former. Therefore, this under-representation may be affecting the selection of descriptors

to represent this type of compounds and their relationship with the property of interest. Removing the two alkynes (hexa-2,4-diyne and hex-1-ynylbenzene) with higher prediction errors, the RMSE decreases around 11.6% to 42.99 and a Q^2 of 0.9684, which is an indicator that these type of compounds are not well represented in the training set. Another class of hydrocarbons with high error rate are the polycyclic compounds, although the experimental confidence on these values is lower than for the rest of the dataset, the fact that they have complex structures and conformations may be the cause for a higher difficulty establishing a relationship between their representation and the property of interest.

Summarizing, the feature selection step yields lower prediction errors (RMSE = 34.10) with a small number of variables (89). When comparing it to using the model with all the available descriptors (1485), the current 89-variable model was able to produce models with an RMSE 23% lower. These reduced errors are relevant in thermochemistry with significant chemical and economical importance.

3.5.2 Case A2 - Predicting Enthalpy of formation and phase change for ThermInfo's dataset

3.5.2.1 Selected chemical descriptors

The list of most important descriptors selected using variable importance calculated by RFs were individually analysed for each property and are made available in Appendix C.2. There are always 30 to 50 variables with substantially higher mean importance and from this point on the importance decreases asymptotically. Figure 3.17 presents the 20 most important variables for each property grouped into general classes of molecular descriptors. In general, it is possible to observe that there are several common descriptors in the top 20 most important variables for each property. Additionally, to predict Standard Molar Enthalpy of Formation (crystalline, gas and liquid phases) the most important variables belong mainly to a class of descriptors that consider the contributions of molar refractivity, partial charges, estate indices, LogP and surface area while to predict Standard Molar Enthalpy of Phase Change (fusion, vaporization and sublimation) the most important variables belong mainly to classes of descriptors that are derived from the constitution and topology of the molecule. It is clear that there

Mean Importance Rank	crys	liq	gas	phasecl	phasecg	phaselg
1	SMR_VSA1	SMR_VSA1	SMR_VSA1	Kappa2	HallKierAlpha	Chi1
2	MaxAbsPartialCharge	MinEStateIndex	PEOE_VSA14	fr_unbrch_alkane	HeavyAtomMolWt	VSA_EState9
3	MinPartialCharge	PEOE_VSA4	MinEStateIndex	PEOE_VSA6	ExactMolWt	MolMR
4	PEOE_VSA1	PEOE_VSA14	VSA_EState8	SlogP_VSA5	MolWt	TPSA
5	fr_NH0	VSA_EState8	SlogP_VSA10	SMR_VSA5	VSA_EState9	Ipc
6	EState_VSA1	fr_alkyl_halide	PEOE_VSA4	EState_VSA5	NumHDonors	LabuteASA
7	fp332	VSA_EState9	VSA_EState9	Chi1n	Chi0	Chi1v
8	SlogP_VSA2	MaxPartialCharge	MaxAbsPartialCharge	Kappa3	BertzCT	HeavyAtomMolWt
9	fr_C_O	FractionCSP3	fr_alkyl_halide	NumRotatableBonds	TPSA	HeavyAtomCount
10	FractionCSP3	HallKierAlpha	MaxAbsEStateIndex	Kappa1	NHOHCount	NHOHCount
11	MaxAbsEStateIndex	MaxEStateIndex	MaxEStateIndex	MolLogP	MolMR	MaxAbsEStateIndex
12	HallKierAlpha	SMR_VSA7	fr_halogen	Chi1v	fr_Al_COO	ExactMolWt
13	MinEStateIndex	MaxAbsEStateIndex	FractionCSP3	EState_VSA4	SlogP_VSA2	NumHDonors
14	MaxEStateIndex	fr_NH0	fp532	MinEStateIndex	SMR_VSA5	MolWt
15	PEOE_VSA14	SlogP_VSA10	MaxPartialCharge	SMR_VSA10	NumRotatableBonds	MaxEStateIndex
16	fp624	SlogP_VSA6	HallKierAlpha	fp439	Kappa1	Chi0
17	NOCount	fr_halogen	Kappa1	PEOE_VSA11	Chi0n	MinPartialCharge
18	TPSA	EState_VSA10	EState_VSA10	fp884	Chi0v	Chi1n
19	fp517	MaxAbsPartialCharge	fp516	BalabanJ	LabuteASA	NumValenceElectrons
20	EState_VSA10	SMR_VSA5	SMR_VSA5	FractionCSP3	Kappa2	BertzCT

Legend:
- Constitutional
- Fingerprints/Fragments
- Topological/Graph Descriptors
- MOE-type descriptors using molar refractivity contributions, partial charges, EState indices, LogP and surface area contributions
- Partial Charges
- Estate indices
- Other/Hybrid

Figure 3.17: List of the 20 most important variables by classes of descriptors for each property (Standard Molar Enthalpy of Formation: crystalline (crys), liquid (liq) and gas phases and Standard Molar Enthalpy of Phase Change: fusion (phasecl), vaporization (phaselg) and sublimation (phasecg)) in case-study A2. More information about the descriptors and their meaning can be found at https://code.google.com/p/rdkit/wiki/DescriptorsInTheRDKit and http://openbabel.org/docs/dev/Fingerprints/intro.html.

are many common descriptors in all physical phases of each group of properties (enthalpy of formation and enthalpy of phase change), especially between the gas and liquid phases.

3.5.2.2 Prediction errors analysis

The prediction errors obtained for each property using the independent validation set were analysed and are represented in the Figure 3.18. Appendix C.2 provides a detailed table of predictive results for the testing set obtained for all properties of case-study A2 using the best model (selected based on training cross-validated results). Similarly to what has already been observed, the model is predicting the enthalpy of formation with small bias, namely the most probable error for **a)**

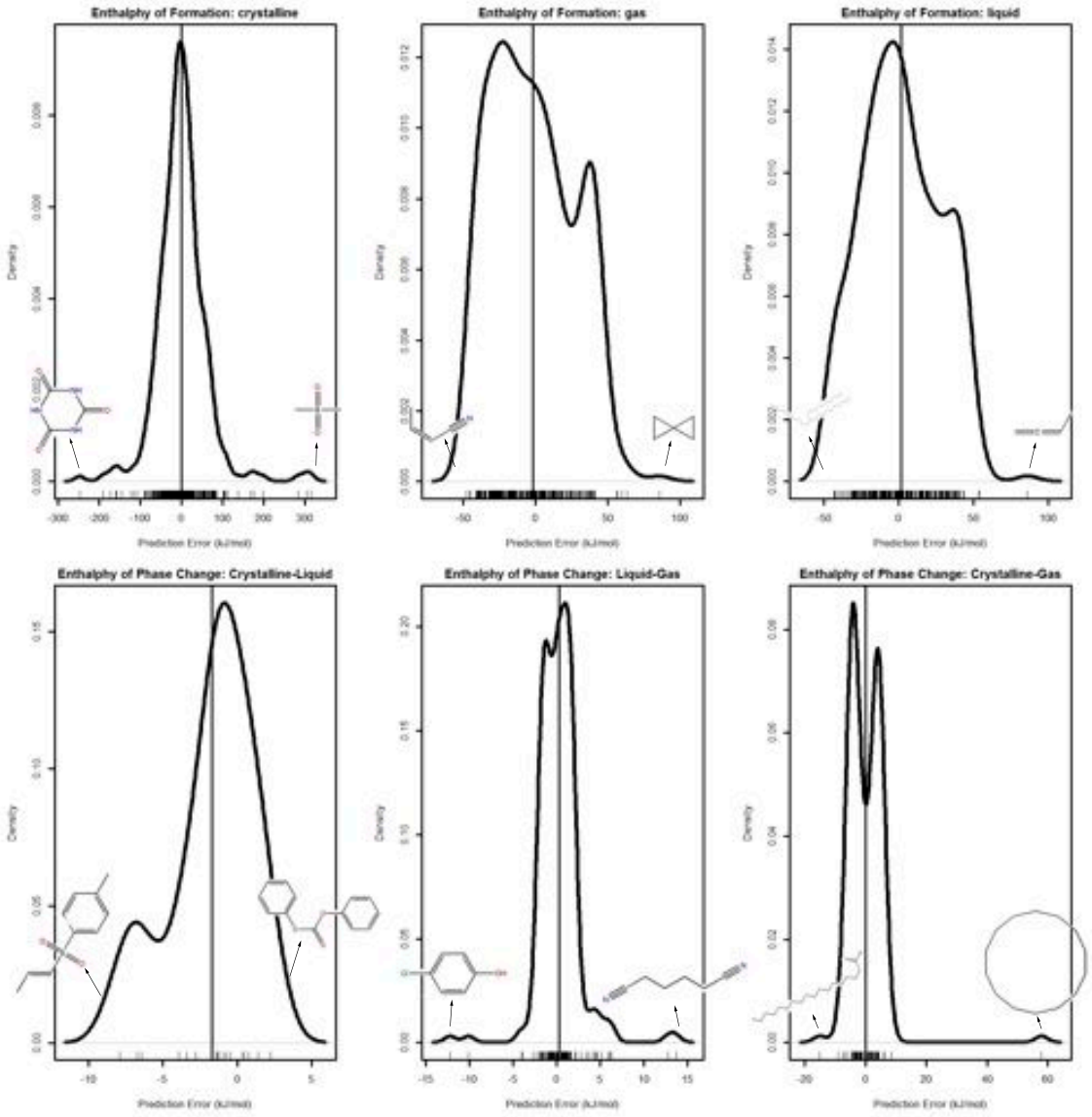


Figure 3.18: Density plots of the differences between the observed values and the predicted values for all properties in the testing sets. The structure of the compounds with most extreme prediction errors are indicated, the positive errors correspond to the compounds Sulphonylbismethane, Spiropentane, 1,2-Butadiene, Carbonic acid diphenyl ester, Hexanedinitrile and Cyclotetradecane (using the order of the plots) and the negative errors correspond to the compounds 1,3,5-Triazine-2,4,6(1H,3H,5H)-trione, (Z)-2-Butenenitrile, 11-Decylheneicosane, (E)-1-Methyl-4-(1-propenylsulphonyl)benzene, 4-Chlorophenol and Octadecanoic acid (using the order of the plots).

crystalline phase is 1.12, **b)** gas phase is -1.76, **c)** liquid phase is 1.60, **d)** fusion is -1.70, **e)** vaporization is 0.25 and **f)** sublimation is 0.05. The compounds with higher errors were also analysed and are represented in Figure 3.18. Once again, the higher errors are mostly compounds with triple bonds or more than one cycle. Small structures with rigid conformations, such as Spiropentane, also showed a higher difficulty establishing a relationship between their representation and the property of interest.

Summarizing, the feature selection step yields lower prediction errors with a smaller number of variables. The number of variables in the model with all the available descriptors (1168) can be reduced by about 80%, increasing the predictive results by 2 to 40%. These reduced errors are relevant in thermochemistry with significant chemical and economical importance.

3.5.3 Case D - NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge

One of the main insights gained during this challenge is that models are as good as the data they are based on, therefore a key limitation to the subsequently use of the produced models is that the set of compounds used to develop the relationship should be similar to those compounds for which predictions are desired. The results obtained for the training set using 10-fold cross validation and test sets are significantly different, however given that the distribution of the test set was biased (as presented in Appendix A in Figure A.7) we cannot conclude about the predictive performance of the produced models, neither if they are applicable to a *real-world* situation. Either the training or test set is not representative of the *real-world* distribution and in such case to develop predictive models, the priors of such distribution in a *real-world* scenario should be known. To test this hypothesis, we merged the training and testing sets of this challenge and randomly sampled it again into training and test set with the same size. The distribution of the median EC_{10} in new datasets is clearly more similar and is presented in Figure 3.19. The RMSE obtained for the new randmly sampled training set is 0.6371 and for the testing set is 0.7146 which are clearly more similar.

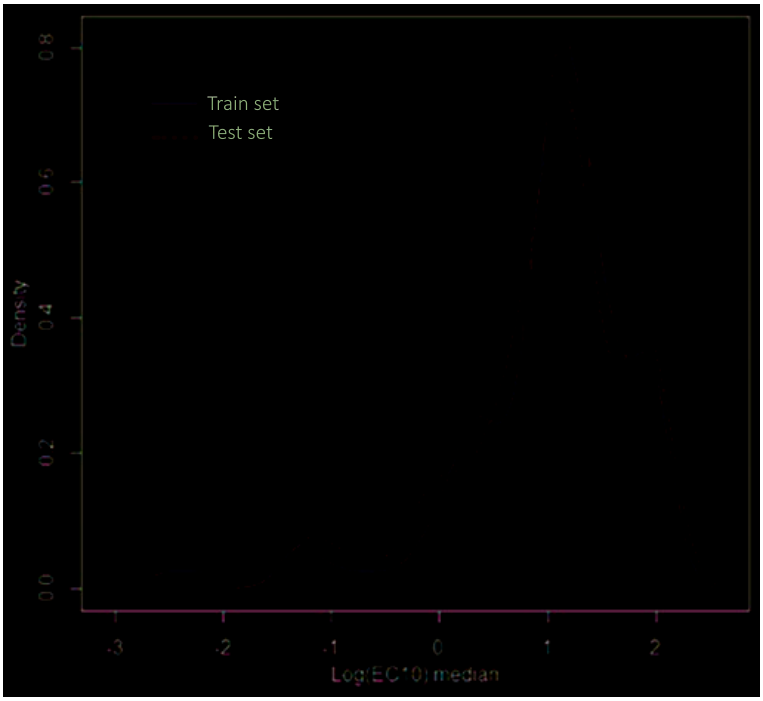
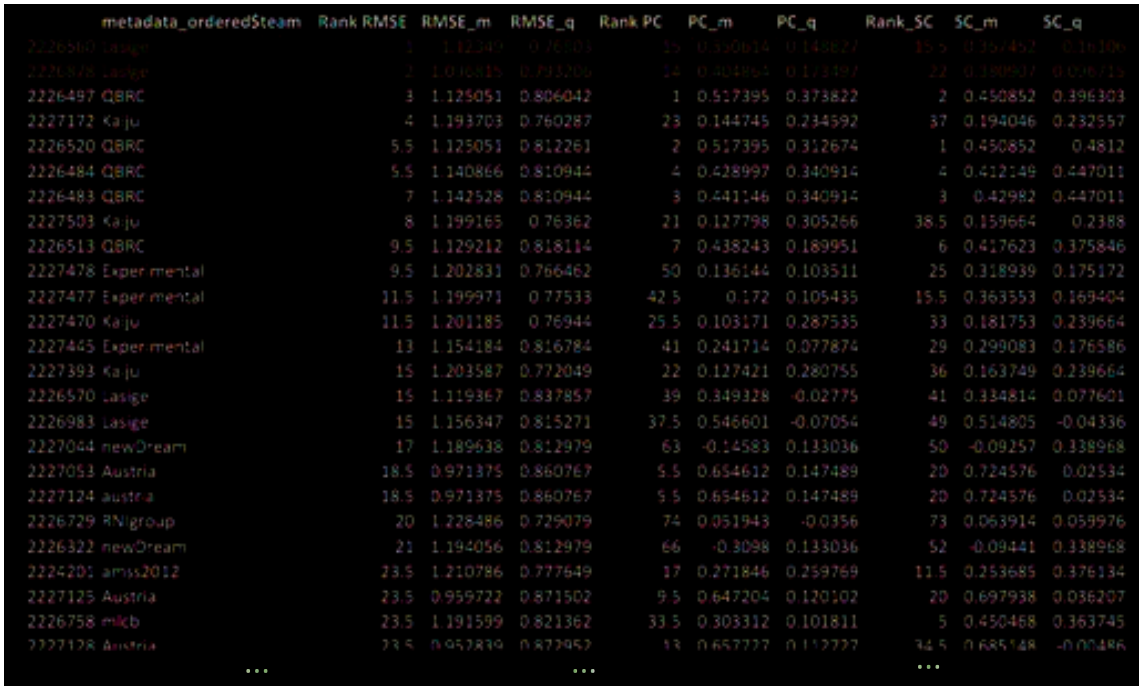


Figure 3.19: Density plot showing the distribution and variation of the median EC_{10} in the train and test sets randomly sampled.

Regarding the objective that we proposed, minimizing the RMSE, the predictive performance of our models ranked position 1 and 2 among 229 participants organized in 24 teams, showing the ability of the model to cope with the defined purpose. However, due to the fact that the testing set and the training set do not show the same statistical properties (meaning they belong to different populations, in the statistical sense), the respective correlation coefficient and the RMSE obtained for both sets appear very weakly correlated. Accordingly, we cannot conclude that the models produced are the best ones to use in a *real-world* scenario for cytotoxicity prediction, even though the approach used was able to give solutions that ranked very competitively among all the other submissions (Figure 3.20).



metadata_orderedSteam	Rank RMSE	RMSE_m	RMSE_q	Rank PC	PC_m	PC_q	Rank_SC	SC_m	SC_q
2226497 QBRC	3	1.125051	0.806042	1	0.517395	0.373822	2	0.450852	0.396303
2227172 Kaju	4	1.193703	0.760287	23	0.144745	0.234592	37	0.194046	0.232557
2226520 QBRC	5.5	1.125051	0.812261	2	0.517395	0.312674	1	0.450852	0.4812
2226484 QBRC	5.5	1.140866	0.810944	4	0.428997	0.340914	4	0.412148	0.447011
2226483 QBRC	7	1.142528	0.810944	3	0.441146	0.340914	3	0.42982	0.447011
2227503 Kaju	8	1.199165	0.76362	21	0.127798	0.305266	38.5	0.159664	0.2388
2226513 QBRC	9.5	1.129212	0.818114	7	0.438243	0.189951	6	0.417623	0.375846
2227478 Experimental	9.5	1.202831	0.766462	50	0.136144	0.103511	25	0.318939	0.175172
2227477 Experimental	11.5	1.199971	0.77533	42.5	0.172	0.105435	15.5	0.363553	0.169404
2227470 Kaju	11.5	1.201185	0.76944	25.5	0.103171	0.287535	33	0.181753	0.239664
2227445 Experimental	13	1.154184	0.816784	41	0.241714	0.077874	29	0.299083	0.176586
2227393 Kaju	15	1.203587	0.772049	22	0.127421	0.280755	36	0.163749	0.239664
2226570 Jasige	15	1.119367	0.837857	39	0.349328	-0.02775	41	0.334814	0.077601
2226983 Jasige	15	1.156347	0.815271	37.5	0.546601	-0.07054	49	0.514805	-0.04336
2227044 newDream	17	1.189638	0.812979	63	-0.14583	0.133036	50	-0.09257	0.338968
2227053 Austria	18.5	0.971375	0.860767	5.5	0.654612	0.147489	20	0.724576	0.02534
2227124 austria	18.5	0.971375	0.860767	5.5	0.654612	0.147489	20	0.724576	0.02534
2226729 RNIgroup	20	1.228486	0.729079	74	0.051943	-0.0356	73	0.063914	0.059976
2226322 newDream	21	1.194056	0.812979	66	-0.3098	0.133036	52	-0.09441	0.338968
2224201 amss2012	23.5	1.210786	0.777649	17	0.271846	0.259769	11.5	0.253685	0.376134
2227125 Austria	23.5	0.959722	0.871502	9.5	0.647204	0.120102	20	0.697938	0.036207
2226758 mlcb	23.5	1.191599	0.821362	33.5	0.303312	0.101811	5	0.450468	0.363745
...				...			...		

Figure 3.20: Final scoring for NIEHS-NCATS-UNC DREAM toxicogenetics challenge. The submissions were evaluated on a final test set of 50 held-out compounds based on the ability of teams to predict the distribution of $log(EC_{10})$ values for each compound in the population, in terms of median $log(EC_{10})$ values and interquantile (q95-q05) distance. The performance of each submission was assessed using Pearson correlation (PC), Spearman correlation (SC) and Root Mean Squared Error (RMSE).

3.5.4 Case G - Blood-Brain Barrier (BBB) Penetration Modelling

The analysis of the model classification results strongly suggests that using a selected smaller number of chemical descriptors is better than using all available information and produces significantly better models. Nonetheless simplistic models with too few descriptors are not enough to produce reliable results. The optimal number of descriptors was found to be around 200, which produced cross-validated results with an expected Mean Squared Contingency Coefficient (MSCC) of 0.687, and overall accuracy of 90.1%, thus definitely superior to the results obtained when using all descriptors: MSCC=0.431 and accuracy = 88.6%.

The list of variable importance in the 10 runs is stable and coherent. The 10 most important variables are consistent with the literature, reflecting mainly 2D information, being the topological polar surface area (TPSA) the most important one. There is a consensus that penetration across the BBB is highly influenced by lipophilicity, which can be quantified using the logP. The results indicate that logP was considered an important descriptor although it occupies the 11^{th} position in the importance list. The use of descriptor set A in conjunction with B and D obtain better results with a smaller number of variables (20), although it is not able to achieve an expected ϕ higher than the one obtained used 200 variables for descriptor sets B and D. This might be related with the fact that the list of most important variables for descriptors sets A + B + D are mostly from set A and represent mostly characteristics of the molecule related with the TPSA, while using descriptors sets B + D alone represent a more diverse set of structural characteristics of the molecule.

In summary, the proposed methodology improves the predictive performance of BBB permeation with 25.6% using just 19% of the original number of descriptors, providing faster and more cost-effective calculation of descriptors by reducing their number, and providing a better understanding of the underlying relationship between the molecular structure and the activity of interest.

3.6 Summary

It is unrealistic to think that all descriptors of a molecule contain useful information for a specific modelling problem. It is further acknowledged that models with larger numbers of variables are not necessarily better. Furthermore, smaller models tend to generalize better than larger models, and tend to be statistically more robust. Therefore, after numerical descriptors have been calculated for each compound, its number should be reduced to a set of them that are information rich while being as small as possible. The proposed approach uses RFs, not as modelling tools for themselves, but as a method capable of identifying the most important features of a given modelling problem, which are then used as input variables to SVM models. It is important to note that RFs were the selected algorithm due to the enumerated advantages; however, in principle, any machine

learning able to produce a ranking of variable importance could be applied. The second part of this hybrid algorithm uses a ranked list of the variables, ranging from the most to the least important, to train SVM models using a stepwise approach of adding one variable for each model according to its predefined rank. Once again, it is important to note that, in principle, any non-linear machine learning method could be applied. The parameters of both models were optimized and the effect of correlated variables studied. From the analysis of the obtained results for different QSPR/QSAR case-studies, we can conclude that the presented methodology performs well for high-dimensional data and it is robust even in the presence of highly correlated variables.

The feature selection step yields lower prediction errors for all case-studies with a smaller number of variables. These reduced errors are relevant with significant chemical and economical importance, but they are also important in terms of computational performance since a smaller number of descriptors need to be calculated producing simpler models that are more robust and comprehensive. It is then safe to conclude that SVMs alone are not able to perform a good optimization, and by combining with a variable selection step we can obtain a minimum subset of important variables to train a faster and more robust model, yielding better prediction performance.

The predictive models were also validated with independent or test sets to assess its performance in new data and the results were similar to the ones obtained for the training set cross validation.

The purpose of the current work was to suggest and apply a methodology able to reduce the variable space while preserving (even increasing) the model prediction capabilities, thus reducing the redundancy and correlation between variables. In summary, this methodology improves the prediction performance of different case-studies using as molecular representation a set of molecular descriptors, providing faster and more cost-effective calculation of descriptors by reducing their number, and providing a better understanding of the underlying relationship between the molecular structure represented by descriptors and the property of interest.

Chapter 4

Representing the Molecular Space Based on Structural Similarity

This chapter describes a fundamental problem in cheminformatics - the definition of a new structural similarity method. This new structural similarity method is the first step in the development of a property prediction approach based on instance-based algorithms. Given that similar molecules tend to have similar physical, chemical and biological properties, the notion of molecular similarity plays an important role in the exploration of molecular datasets, query-retrieval in molecular databases, and in structure-property/activity modelling. Various methods to define structural similarity between molecules are available in the literature, but so far none have been used with consistent and reliable results for all situations.

At the center of the methodology developed to define structural similarity is the concept of atom alignment. This method is based on the comparison of bonding profiles of atoms on comparable molecules, including features that are seldom found in other structural or graph matching approaches like chirality or double bond stereoisomerism. The similarity measure is then defined on the annotated molecular graph, based on an iterative directed graph similarity procedure and optimal atom alignment using a pairwise matching algorithm.

This non-contiguous atom matching structural similarity method (NAMS) was tested and compared with one of the most widely used similarity methods (Fingerprint-based similarity) using three difficult case-studies (Cases **A1**, **E** and **F**) described in detail in Appendix A.

4.1 Measuring Molecular Similarity

Beauty is in the eye of the beholder.
~ Margaret Wolfe Hungerford, Molly Bawn (1878)

Molecules are typical examples of unstructured data for which tasks such as searching, sorting, analysing and extracting knowledge are challenging. A molecule can have an arbitrary dimension, structure and composition, and moreover, there is not an univocal and unequivocal way of coding and comparing these molecules. Several computational tools have been developed over the years in pursuance of solving this issue. Fundamental observations that justify the amount of methods developed to compare molecules derive from the fact that similarity has a context (Bender & Glen, 2004) and the representation of molecular structures implies information loss. Researchers have explored the concept of similarity between molecules which provides an important approach to search databases, predict properties of compounds, design structures with a predefined set of properties and conduct structure-based drug design studies (Auer & Bajorath, 2008; Bajorath, 2001; Bender & Glen, 2004; Eckert & Bajorath, 2007; Johnson & Maggiora, 1990; Kubinyi, 1998; Nikolova & Jaworska, 2003; Sheridan & Kearsley, 2002; Willett, 2005). These studies are based on the "neighbourhood" premise, which states that similar molecules usually have similar activities and properties (Bender & Glen, 2004; Johnson & Maggiora, 1990; Patterson *et al.*, 1996). The definition of similarity between molecules consists of comparing chemical structures, specifically representing these molecules and quantifying the similarity between them. Various methods to define structural similarity between molecules are available in the literature (Bender & Glen, 2004; Nikolova & Jaworska, 2003). The most popular approaches to represent the structure of molecules under comparison can be divided in three broad categories: approaches based on structural

descriptors (two- and three-dimensional), molecular fragments and graph matching (descriptor-independent methods).

4.1.1 Approaches based on structural descriptors

Methods based on structural descriptors attempt to describe the information encoded in the molecular structure into a set of numerical values and define some means for comparing them (Nikolova & Jaworska, 2003). As already described in chapter 2, a large number of different descriptors can be used in similarity calculations and they differ in the complexity of the encoded information and in the computation time (Todeschini & Consonni, 2009). A descriptor positions each abstract molecular representation in the descriptor space. It is then possible to compare molecules, considering that the distance of the abstract molecular representations reflects their similarity in this specific descriptor space (Bender & Glen, 2004). Depending on the context of the comparison the appropriate set of descriptors may change. For instance, in the property prediction context using a certain descriptor space, a set of structurally similar molecules may be also similar with respect to one property A, but completely dissimilar with respect to a property B (Bender & Glen, 2004). Molecular similarity is a non-linear problem for which there is not a set of descriptors or a similarity measure that correlates with every context of comparisons one can perform (Bender *et al.*, 2009; Kubinyi, 1998). Moreover, published works have shown that compounds that are similar to known active molecules are themselves far less frequently active than one might expect (Martin *et al.*, 2002; Nikolova & Jaworska, 2003). Three-dimensional descriptors were expected to have a better performance than two-dimensional descriptors, however, the opposite trend has been verified in several studies (Hu *et al.*, 2012; Sheridan & Kearsley, 2002).

4.1.2 Approaches based on molecular fragments

Molecular fragments are structural descriptors which indicate the presence of some structural fragments in molecules. An important class of molecular fragment descriptors are fingerprints which have been presented in great detail in chapter 2. Fingerprints are one of the most widely used methodologies for global

molecular similarity analysis, despite the fact that this molecular representation has some disadvantages such as information loss and a bias in the evaluation of molecular similarity due to differences in molecular complexity and size. There is information loss when representing molecules as fingerprints since, for example, binary fingerprints simply indicate the presence or absence of a given fragment rather than set bits for the number of matches and the library of structural fragment fingerprints may not include fragments that are important for certain problems. Figure 4.1 exemplifies the problems mentioned above with four compounds with the same fingerprint yet totally different values of a physico-chemical property.

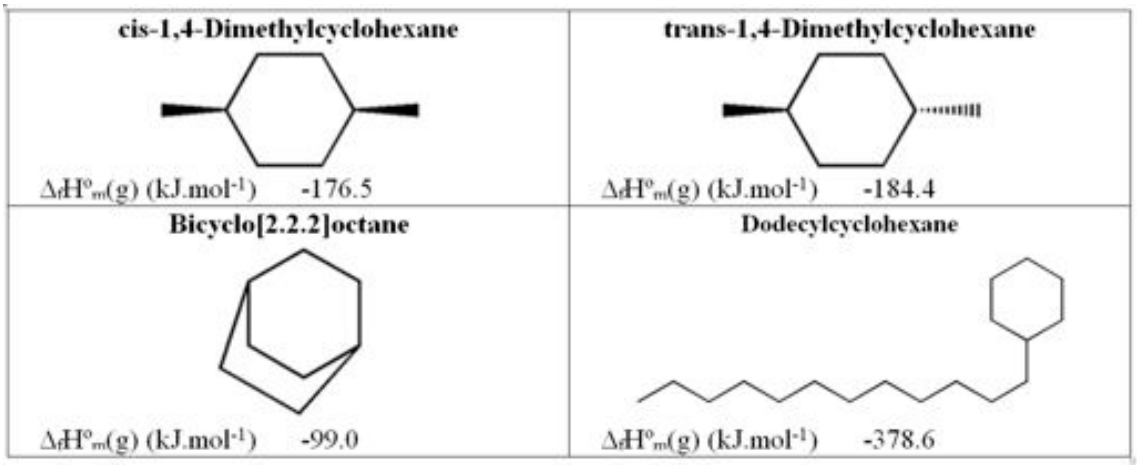


Figure 4.1: Example of four compounds (cis-1,4-Dimethylcyclohexane., trans-1,4-Dimethylcyclohexane, Bicyclo[2.2.2]octane, Dodecylcyclohexane) with the same fingerprint (100% similarity according to Tanimoto coefficient) and respective value of enthalpy of formation in gas phase.

Also, the average similarity appears to increase with the complexity and size of the query compound, since there is an higher bit density than for simpler molecules which will endorse a larger overlapping of the fingerprints (Bender & Glen, 2004; Eckert & Bajorath, 2007; Flower, 1998; Holliday *et al.*, 2002; Tovar *et al.*, 2007). Fingerprints can also be used to represent molecules in the context of property prediction or to efficiently filter out dissimilar structures from a dataset, since quantifying two fingerprints as very dissimilar means, in principle, that the underlying structures are certainly dissimilar (Flower, 1998).

4.1.3 Approaches based on graph matching

A molecule can also be represented, using graph theory, as a labelled graph whose vertices correspond to the atoms and edges correspond to the covalent bonds. The representation of molecules using graphs has some advantages, namely, graphs are intuitive when representing a molecule since they are close to our understanding of a molecule and they have a solid mathematical background with different existing techniques to compare labelled graphs (Ehrlich & Rarey, 2011). However, representing molecules as graphs raises an important issue, identical graphs do not necessarily represent identical structures and *vice-versa*. This problem is originated by mesomeric structures (e.g. aromatic rings), stereochemistry (e.g. chirality), tautomeric forms, among others (Ehrlich & Rarey, 2011).

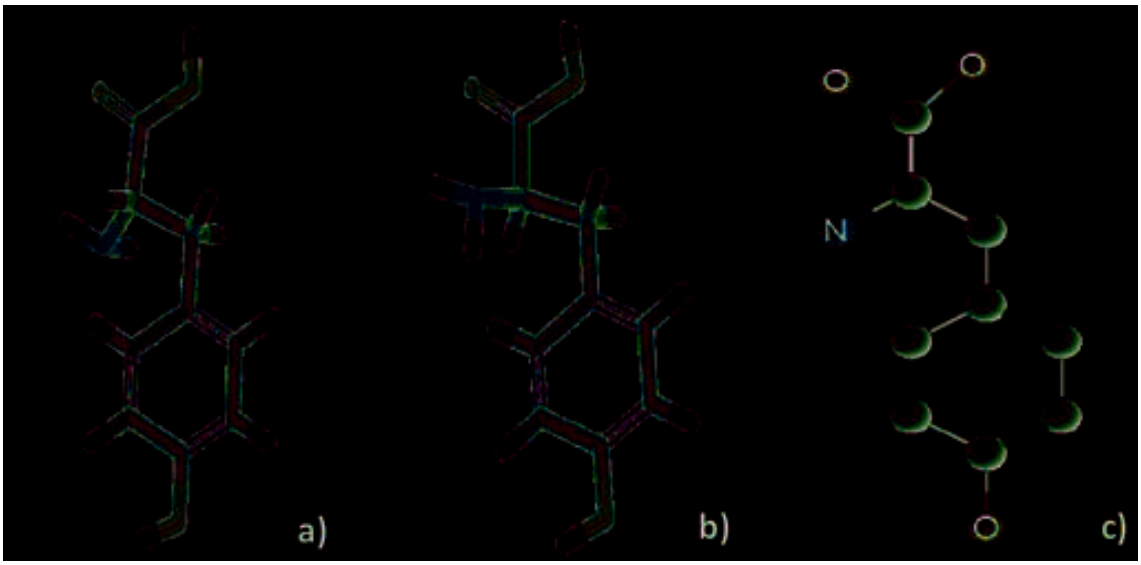


Figure 4.2: 3-D representation of a) (R)-2-amino-3-(4-hydroxyphenyl) propanoic acid and b) (S)-2-amino-3-(4-hydroxyphenyl) propanoic acid. c) molecular graph representation of 2-amino-3-(4-hydroxyphenyl) propanoic acid (excluding hydrogen atoms).

Figure 4.2 exemplifies two different compounds, a) (R)-2-amino-3-(4-hydroxyphenyl) propanoic acid and b) (S)-2-amino-3-(4-hydroxyphenyl) propanoic acid with identical connectivity (Figure 4.2 - c), i.e. a) and b) have the same topology, but different topography.

The graph matching approach is descriptor-independent and often employs the concept of the maximum common sub-graph (MCS) (Barnard, 1993; Ehrlich & Rarey, 2011; Raymond & Willett, 2002). A common substructure can be defined as a substructure present in two molecules with the same bonding profile and therefore the objective of a MCS algorithm is to find the common substructures with the largest number of atoms and bonds (Ehrlich & Rarey, 2011). These searches tend to be time-consuming due to the NP-complete complexity of the sub-graph isomorphism problem (a NP-complete problem cannot be solved in polynomial time) (Garey & Johnson, 1979), however many approximated heuristics have been proposed to overcome this complexity. These heuristics are based on techniques such as pruning the search tree of the exact algorithm (Rahman *et al.*, 2009; Raymond & Willett, 2002), greedy algorithms (Berglund & Head, 2010; Hagadone, 1992; Kawabata, 2011), genetic algorithms (Brown *et al.*, 1994; Wang & Zhou, 1997), reduced representations of chemical graphs (Batista *et al.*, 2006; Rarey & Dixon, 1998; Takahashi *et al.*, 1992), among others. This group of descriptor-independent methods benefits from improved sensitivity in relation to descriptor/fragment based similarity searches since they can find atom-atom, bond-bond or atom-bond equivalences between query and target molecules (Ehrlich & Rarey, 2011; Garcia *et al.*, 2003).

4.1.4 Quantifying the degree of similarity/dissimilarity between molecules

Following the selection of a molecular representation, to determine the numerical value of the similarity/dissimilarity of the molecules it is necessary to compare their abstract representation using a similarity coefficient/distance measure. This quantification can be obtained using simple distance measures such as Hamming or Euclidean, and association/similarity coefficients such as Tanimoto-Jaccard, Dice or Cosine (Nikolova & Jaworska, 2003; Willett *et al.*, 1998). Distance measures consider a shared absence of fragments as evidence of similarity while association coefficients consider a shared presence of fragments as evidence of similarity, ignoring molecular features that are absent in both molecules. While the Tanimoto-Jaccard coefficient is the most popular similarity coefficient, there are

several others that have been explored (Gillet *et al.*, 1998; Holliday *et al.*, 2002). The question then is which coefficient performs better given the selected molecular representation.

Similarity is an abstract, problem-dependent and subjective concept and its definition is, to a great extent, a semantic question. Similarity depends on comparative perception without a defined standard, in a certain degree "like beauty, it is in the eye of the beholder" (Bajorath, 2004). When judging, for example, the similarity of faces, some may consider that two faces are similar if they have a common complexion, while others would consider other facial characteristics such as the eyes, the nose, the ears and the mouth. Because of this subjectivity it is difficult to develop methods for unambiguously quantifying the similarities of objects such as molecules. Moreover, in many situations, while two molecules are not similar, some of their parts are, the challenge is then the quantification of the degree of partial similarity between the given molecules. Some authors argue that all pattern recognition problems boil down to giving a quantitative interpretation of similarity between objects (Bronstein *et al.*, 2008). There are also studies that show that using the existing methods to quantify similarity it is not always possible to take advantage of the similarity principle to predict properties/activities of molecular structures with a good performance (Kubinyi, 1998; Martin *et al.*, 2002).

4.1.5 Non-contiguous atom matching structural similarity

A new atom alignment method for adequately quantifying the structural similarity between molecules with an high discriminative power of similar molecules was developed (Teixeira & Falcao, 2013). In general, to solve the global problem of quantifying the structural similarity between molecules, we decided to break it down into solvable different parts by reducing the molecule to atoms and compare atoms of different molecules in order to find the best alignment between them. These atoms should be considered not only by their intrinsic chemical characteristics but also according to their relation to the other atoms in the molecule. The similarities detected by an atom correspondence approach like the present

one are consistent with the chemistry and structure of the molecules because it depends on the direct neighbourhood of each atom as well as the overall topology of the molecule, becoming more intuitively understood because similar atoms in the molecules are explicitly shown. The relation between each atom and the whole molecule allows the consideration of important characteristics of the atoms and bonds such as the chirality and the double bond stereoisomerism, since these depend on the orientation and symmetry of the neighbouring atoms in the space (Teixeira *et al.*, 2013a). Although these characteristics are ignored by most 2D similarity methods, they are of great importance in many different fields since the molecular properties and biological effects of the stereoisomers are often significantly different (Islam *et al.*, 1997).

For the comparison of the bonding profiles of atoms on comparable molecules, we defined three main steps. Firstly, a set of attributes should be selected in order to characterizes the molecule's atoms (e.g. atom type and chirality) and bonds (e.g. bond type and stereoisomerism) and all the topological relations between the atoms for the purpose of their comparison. Secondly, an adjustable weighting scheme that emphasizes certain characteristics and account for the differences with a penalty function, in accordance with the context of the problem. Thirdly, determining a value for a measure that represents the degree of similarity between the annotated molecular graph, based on a recursive concept of graph similarity and an optimal alignment between atoms using an heuristic approach.

In the following sections this Non-contiguous Atom Matching structural Similarity (NAMS) algorithm will be presented in detail, including the atom and bond matching functions, the alignment between the pair of molecules under comparison and the quantification of the resemblance of the molecules. To clarify the application of the method, an illustrative example is also presented and to demonstrate the granularity and effectiveness of this method, we present similarity analyses for three different case-studies (Cases **A1**, **E** and **F**) described in detail in Appendix A and compare the results against a similarity function based on path-based fingerprints.

4.2 Development of the Non-contiguous Atom Matching Structural Similarity

There is in my opinion a great similarity between the problems provided by the mysterious behaviour of the atom and those provided by the present economic paradoxes confronting the world.

~ Paul Dirac, Nobel Prize in Physics (1933)

4.2.1 Concepts and overview

NAMS is based on the concept of atom matching between two molecules, that is, for every atom of a molecule A find the atom of a molecule B that is more related to it, by scoring every possible atom comparison between A and B. If it is possible to define a score for each atom of A as related to B, it is then possible to get the best possible matchings by selecting the atom matchings that produce the best possible score, with the constraints that one atom of A may only be matched to one atom of B and vice-versa. The main issue is then the definition of an atom scoring function. The present approach is non-contiguous as it may happen that matchings between atoms of different molecules may not reflect the contiguous atomic fragments of the other molecule.

Atoms are not isolated objects within molecules, their characteristics depend on a) the bonds to other atoms and b) the neighbouring atoms. Yet the characteristics of these neighbouring atoms depend as well on their bonds and their own neighbouring atoms, and accordingly until all molecular bonds and atoms are exhausted within the molecule. The main idea is that each atom (α_{Ai}) of a given molecule A at position i can be represented by a corresponding graph that is centred in α_{Ai} and encompasses all the other bonds and atoms in the molecule. As such, the procedure for comparing atoms is essentially a procedure for comparing graphs, where each graph is a *view* of the full molecule from that atom. These graphs are directed graphs as there are different molecular graphs for each atom

However, the problem of comparing directed graphs is computationally expensive (Garey & Johnson, 1979; Kobler *et al.*, 1993) even for moderately sized instances. Therefore, the following heuristic procedure was devised. Primarily,

a simplified representation of the atomic directed graph was adopted. This representation encompasses a list of all the bonds of the molecule coupled to their topological distances to the atom α_{Ai} under observation. Each bond is represented by the start and end atoms as well as the covalent chemical bond between them, that is an *atom-bond-atom* tuple that, through the document, will be designated as an *aba-bond*). Secondly, using this simplified representation, the procedure of comparing different atoms of different molecules becomes a localized problem of trying to match the best possible representation of an atom as related to its molecule, by matching each *aba-bond* pairings depending on their topological distances to the atoms being compared. The bond matching problem can be solved using a distance function and the same assignment algorithm suggested for atom matching. Finally, using the best possible alignment between the molecules under comparison as determined by the atom and bond matching functions, it is possible to produce a score that indicates the degree of superimposition between these molecules.

The procedure described above can then be outlined in the following algorithm:

1. For each bond of each molecule discriminate each *aba-bond* of the molecule, their structural characteristics and their respective topological distances as related to each atom in the molecule;

2. Use a distance-dependent bond matching function to compute a matching score between each *aba-bond* of each atom of each molecule and produce, for each two atoms being compared, a bond matching matrix;

3. Use an assignment algorithm to compute the best possible matching score between each possible pair of atoms, by matching the *aba-bonds* matrices of each atom being compared. The resulting matrix will have a score that quantifies how closely each atom of molecule A matches any other atom of molecule B;

4. Use the same assignment algorithm to assign each atom of molecule A to each atom of molecule B;

5. The similarity score between molecules A and B is then the sum of the similarities between the best possible atom alignments;

6. Compute a similarity coefficient by calculating the ratio between the molecule A and B superimposition and the sum of the self-superimposition of molecules A and B.

An important step of the algorithm is the *aba-bonds* comparison, as for each bond in the molecule, the start and end atoms are accounted, as well as different structural characteristics of the atoms and bonds. In the implementation of the algorithm the following structural characteristics of the atoms and bonds can be included, the nature of the atomic elements with distinct atomic similarity functions as well as the nature of the bond (single, double, triple, aromatic), chain type (e.g. linear or cyclic), include/exclude hydrogen atoms and even include other specific characteristics of bonds or atoms that are dependent on the topology and geometry of the neighbourhood, namely atomic chirality and cis-trans bond isomerism (Teixeira *et al.*, 2013a). However, other structural characteristics of the atoms and bonds can be easily included.

In the following sections each step of the algorithm is detailed, clarifying the implementation decisions and illustrating with a simple example for comparing two small molecules. For presenting each part of the method a bottom-up approach will be followed, first describing the *aba-bond* matching procedures, then the atom matching using the scores produced for each atom, and finally the definition of a molecular similarity score.

4.2.2 Molecular Alignment by Bond Matching

Two desirable characteristics of a bond similarity function should be: (1) it produces a score based on bond characteristics, and (2) it includes a factor that, for the atoms under comparison, accounts for the respective topological distances. These characteristics and how the method makes use of them to produce a molecular alignment by bond matching will be described in the following sections.

4.2.2.1 Bond Similarity

A chemical bond matching function should involve the computation of a similarity function between the pre-defined set of characteristics of any two bonds. For directly matching *aba-bonds* a product function that multiplies the similarities between the set of characteristics of the bond that includes not only the bond itself, but also both end atoms was devised. By using a product function, each difference in the bond characteristics cumulatively decreases the end result, which asymptotically approaches 0, thus effectively working as a similarity function. Therefore if a generic *aba-bond* β can be defined by a tuple of P characteristics (h_m): $\beta = (h_1, h_2, ..., h_P)$, a way to compute a similarity function between bonds β^k and β^l is by cumulatively multiplying their paired attributes' similarity:

$$V_{nd}(\beta^k, \beta^l) = \prod_{m=1}^{P} W_m(h_m^k, h_m^l) \tag{4.1}$$

Where W_m is a function that outputs a value from 0 to 1.0, when comparing the same set of characteristics for two bonds. A resulting $W_m = 1.0$, means that the characteristics are exactly the same, whereas a value of $W_m < 1.0$ implies a difference.

4.2.2.2 *Aba-bond* distance-compensation functions

The inclusion of the bond topological distances to atoms is fundamental for the bond matching system. Yet two characteristics are deemed important for adequately using topological distances to weight *aba-bond* matching: a) *aba-bonds* that are closer to the atoms being matched should have a larger impact on the matching score than *aba-bonds* that are very far topologically from the atoms being compared; b) it should be possible to pair two *aba-bonds* even if they appear in distinct levels, but such pairing should have a lower score in the final matching function. Several possibilities for such functions exist, but the following empirical and parametrized *aba-bond* distance compensation function that respects these requirements was devised:

$$V(\beta_{Ai}^k, \beta_{Bj}^l) = \frac{V_{nd}(\beta_{Ai}^k, \beta_{Bj}^l)}{(|d_{Ai}^k - d_{Bj}^l| + \max(d_{Ai}^k, d_{Bj}^l) + 1.0))^{\mu}} \tag{4.2}$$

Where d^k_{Ai} and d^l_{Bj} are the distances of *aba-bonds* β^k and β^l to atoms α_{Ai} and α_{Bj} respectively. The parameter μ was used as a way to weight the importance of the bond distance to an atom, when $\mu \geq 0$. Lower values for the μ parameter disregard the importance of distance of the *aba-bonds* to the atoms being compared, thus focusing on a more functional and local matching, whereas larger values emphasize the importance of distance and add a more global matching of the structure pendant to the score.

4.2.2.3 *Aba-Bond* Matching Function

As referred, all atoms within a molecule must be evaluated in relation to all the other atoms in that same molecule. Therefore, to compare two atoms of different molecules these relationships must be taken into account. Consequently, all the molecule bonds and their relations to the atom α_{Ai} are compared to all the bonds of the other molecule as related to the atom α_{Bj} with the constraint that at most one bond of a molecule can be associated to one bond of the other molecule. Thus,

$$T(\alpha_{Ai}, \alpha_{Bj}) = \max \sum_{i=0}^{M_A} \sum_{j=0}^{M_B} b_{kl} V(\beta^k_{Ai}, \beta^l_{Bj}) \tag{4.3}$$

such that,

$$\sum_{k=1}^{M_A} b_{kl} <= 1 \quad \forall j \leq M_B \tag{4.4}$$

$$\sum_{l=1}^{M_B} b_{kl} <= 1 \quad \forall i \leq M_A \tag{4.5}$$

Where M_A and M_B are the number of bonds present in molecules A and B respectively, b_{kl} is a parameter set to 1 if the bond β^k as related to the atom α_{Ai} (β^k_{Ai}) is matched to the bond β^l as related to the atom α_{Bj} (β^l_{Bj}) and 0 otherwise. Function $V(\beta^k_{Ai}, \beta^l_{Bj})$ represents the similarity between bonds weighted by the respective topological distances of the *aba-bond* as described above (Equations (4.1) and (4.2)). Equations (4.3) to (4.5) present a optimization problem

that can be solved using the well-known and the most efficient algorithm for solving the assignment problem - Kuhn-Munkres Algorithm (Kuhn, 1955; Munkres, 1957), of known complexity $O(n^3)$ where $n = \max(M_A, M_B)$, for which there is a guaranteed optimal solution within polynomial time. This algorithm models an assignment problem as a non-negative $N \times M$ similarity matrix, where each element represents the similarity of assigning the *ith aba-bond* of an atom in Molecule A to the *jth aba-bond* of an atom in Molecule B, and it figures out the solution that maximizes the similarity, choosing a single item from each row and column in the matrix, such that no row and no column are used more than once. The method is based on the principle that if a constant is subtracted to the elements of similarity matrix, the optimum solution of the assignment problem is the same as the original problem. Original similarity matrix is reduced to another similarity matrix by subtracting the row minimum value from each row and the column minimum value from each column and this process is repeated until all possible rows and columns have an assigned value of zero.

4.2.3 Molecular Alignment by Atom Matching

For any two molecules A and B, a global atomic matching is the best possible matching of all atoms of molecule A to all atoms of molecule B. Formally, for each atom α_{Ai} of molecule A, the purpose is to find the best matching atom α_{Bj} of molecule B. This approach requires the *aba-bond* matching function $T(\alpha_{Ai}, \alpha_{Bj})$ presented above (Equations (4.3) to (4.5)) to compute a matching score between any two atoms (α_{Ai} and α_{Bj}). Therefore, for two molecules A and B with a total number of atoms N_A and N_B respectively, the goal is to find the similarity ($S(A, B)$) between them, which represents the optimal matching between their atoms:

$$S(A, B) = \max \sum_{i=0}^{N_A} \sum_{j=0}^{N_B} a_{ij} T(\alpha_{Ai}, \alpha_{Bj}) \tag{4.6}$$

such that,

$$\sum_{i=1}^{N_A} a_{ij} <= 1 \quad \forall j \leq N_B \tag{4.7}$$

$$\sum_{j=1}^{N_B} a_{ij} <= 1 \quad \forall i \leq N_A \tag{4.8}$$

Where a_{ij} is a binary variable, set to 1 if atom i that belongs to molecule A (α_{Ai}) is matched to atom j belonging to molecule B (α_{Bj}) and 0 otherwise. The constraints (4.7) and (4.8) ensure that at most one atom of molecule A is matched against at most one atom of B and, respectively, that at most one atom of B is matched against at most one atom of A. The problem is then to find the set of values for a_{ij} that maximize the similarity score $S(A, B)$. With this approach, the subsequent formulation is once again a typical assignment problem that can be solved in polynomial time using the Kuhn-Munkres Algorithm (Kuhn, 1955; Munkres, 1957), as described above for *aba-bond* matching, this time, however, for atom matching.

4.2.4 Translating the Molecular Alignment into a Structural Similarity Score

At this stage, the atoms from both molecules were matched and their optimal alignment scored. As the value computed measures the intersection (superimposition) between the molecules A and B ($S(A, B) = A \cap B$), it is possible to compute the similarity score between both molecules by using the self-scores produced by self-superimposition ($S(A, A)$ and $S(B, B)$). Other similarity measures could be readily applied, however Jaccard-Tanimoto coefficient (JC) it is not only appropriate but also by far the most widely used. This similarity measure represents the fraction of the common parts of both molecules relative to the union of all their parts:

$$JC_{AB} = \frac{A \cap B}{A \cup B} = \frac{S(A, B)}{S(A, A) + S(B, B) - S(A, B)} \tag{4.9}$$

Where $S(A, A)$ and $S(B, B)$ is the self-superimposition of molecules A and B, respectively. JC_{AB} is a similarity coefficient where a $JC_{AB} = 1.0$ indicates total superimposition between both molecules, and thus identity, while a $JC_{AB} = 0.0$ shows no point of intersection between both molecules.

4.2.5 An illustrative example of the application of NAMS

One simple example of the application of the presented method will help in clarifying its application to a concrete case. Two small molecules will be compared, namely Molecule *A*: *3-methylbut-3-en-2-amine* (SMILES: CC(=C)C(N)C) and molecule *B*: *cyclopropenyl methanamine* (SMILES: NCC1=CC1) (Figure 4.3).

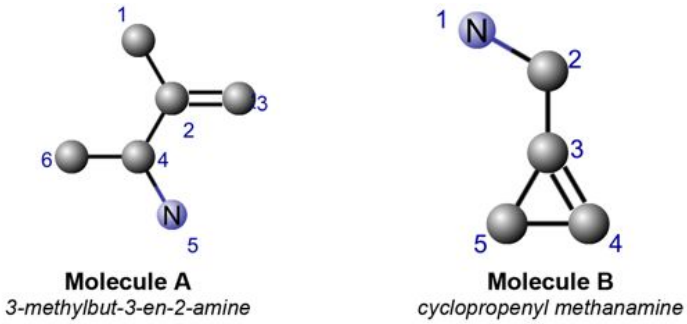


Figure 4.3: Chemical structures of the example molecules **(A)** *3-methylbut-3-en-2-amine* (SMILES: CC(=C)C(N)C) and **(B)** *cyclopropenyl methanamine* (SMILES: NCC1=CC1). The canonical numbers of each atom are indicated near each atom symbol.

For simplification, only three structural characteristics of the *aba-bond* will be considered (h_m): left end-atom element nature, right end-atom element nature and bond order. Considering then four different *aba-bonds* that occur in molecule *A*: $(N-C)$, $(C-N)$, $(C-C)$ and $(C=C)$. For this example, the respective W_m functions (degree of similarity between the set of characteristics of any two bonds) will be defined with very simple rules:

$$W_1 = \begin{cases} 0.1 & \text{if atom elements on left end of the bonds are different} \\ 1.0 & \text{if atom elements on left end of the bonds are equal} \end{cases} \tag{4.10}$$

$$W_2 = \begin{cases} 0.1 & \text{if atom elements on right end of the bonds are different} \\ 1.0 & \text{if atom elements on right end of the bonds are equal} \end{cases} \tag{4.11}$$

$$W_3 = \begin{cases} 0.8 & \text{if covalent bond orders are different} \\ 1.0 & \text{if covalent bond orders are the same} \end{cases} \tag{4.12}$$

Table 4.1: Sample similarity values (V_{nd}) between *aba-bonds* extant in Molecule A (*3-methylbut-3-en-2-amine*).

V_{nd}	$N-C$	$C-N$	$C-C$	$C=C$
$N-C$	1.00	0.01	0.10	0.08
$C-N$	0.01	1.00	0.10	0.08
$C-C$	0.10	0.10	1.00	0.80
$C=C$	0.08	0.08	0.80	1.00

With these rules, equation (4.1) is then used for computing the *aba-bond* topological distances. As a concrete example, $V_{nd}((N-C),(C=C))$ is calculated by iteratively applying the rules: W_1 - as the left end atom elements of each *aba-bond* differs ($Nitrogen \neq Carbon$) apply a factor of 0.1 (rule (4.10)); W_2 - the right end atom elements are the same ($Carbon = Carbon$) so the factor to apply is 1.0 (rule (4.11)); W_3 - as the bond orders (single covalent bond versus double covalent bond) are different so the factor to apply is 0.8 (rule (4.12)). For these two *aba-bonds*, $V_{nd}((N-C),(C=C)) = 0.1 \times 0.8 \times 1.0 = 0.08$. Table (4.1) displays the similarity values between the four types of *aba-bonds* extant in Molecule A.

These values are the results of the comparison of different *aba-bonds* irrespective of their topological distances to each atom. To account for the topological distances for each calculated V_{nd} it is necessary to check each *aba-bond* as relative to each atom of the molecule. For the molecule A (Figure 4.3), starting at each of the atoms of the molecule, all possible *aba-bond* topological levels are displayed on Table (4.2).

With the bond similarities determined, by calculating the $V_{nd}(\beta^k, \beta^l)$ coefficients using Equation (4.1), it is then possible to compute the similarities between any two atoms, by applying the *aba-bond* distance-compensation function (Equation (4.2)), which weights, as detailed before, each bond similarity of two different atoms according to their topological distance. The equation (4.2) requires one user-defined parameter (μ) that only affects the denominator, which is also dependent on the relative topological distances of each *aba-bond* to the respective reference atom. As an example, the atom C_1 will be compared with the atom C_6 within the molecule A, where each of them has only 3 distinct topological

Table 4.2: *Aba-Bonds* topological distances (d) starting from each atom α_i (where α is the atomic symbol of the atom in the position i as represented in Figure 4.3) of the molecule A (*3-methylbut-3-en-2-amine*).

Atom α_i	**Topological Distance**		
	$d = 0$	$d = 1$	$d = 2$
C_1	$C_1 - C_2$	$C_2 = C_3$ $C_2 - C_4$	$C_4 - N_5$ $C_4 - C_6$
C_2	$C_2 - C_1$ $C_2 = C_3$ $C_2 - C_4$	$C_4 - N_5$ $C_4 - C_6$	
C_3	$C_3 = C_2$	$C_2 - C_1$ $C_2 - C_4$	$C_4 - N_5$ $C_4 - C_6$
C_4	$C_4 - C_2$ $C_4 - N_5$ $C_4 - C_6$	$C_2 - C_1$ $C_2 - C_3$	
N_5	$N_5 - C_4$	$C_4 - C_6$ $C_4 - C_2$	$C_2 = C_3$ $C_2 - C_1$
C_6	$C_6 - C_4$	$C_4 - C_2$ $C_4 - N_5$	$C_2 = C_3$ $C_2 - C_1$

levels (d = 0, 1 and 2) as represented in Table 4.2. Assuming that $\mu = 2.0$, the denominators of Equation (4.2) can be pre-calculated as presented in Table 4.3. Table 4.3 shows that for comparing *aba-bonds*, the smaller denominators are, as expected, the ones for *aba-bonds* that are closer to each other, producing higher *aba-bond* similarity scores, while there is an increase of the denominator as the topological distance between the *aba-bonds* grows, producing smaller *aba-bond* similarity scores. For instance, when comparing two *aba-bonds* that are both at $d = 1$, the similarity score of the bond decreases to 1/4, whereas if one bond is at $d = 0$ and the other at $d = 1$, the distance weighting factor is 1/9.

Equation (4.2) can now be applied to calculate the similarities between all possible bond matchings of any 2 given atoms, using the *aba-bond* similarities presented in Table 4.1 as numerator and the distance coefficients presented in Table 4.3 as denominator. Table 4.4 presents the results of *aba-bond* similarity matching between the atoms C_1 and C_6 of molecule A weighted by their topological distance. To finally compute the similarity between these 2 atoms,

Table 4.3: Denominators of the Equation (4.2) for bond comparison based on the topological distances d_1^k and d_6^l of the *aba-bonds* β^k and β^l to atoms C_1 and C_6 of Molecule A, considering $\mu = 2.0$

d_1^k \ d_6^l	0	1	2
0	1	9	25
1	9	4	16
2	25	16	9

Table 4.4: *Aba-bond* scores as calculated by Equation (4.2) for the atoms C_1 and C_6 of molecule A, considering their topological distances ($d(C_1, aba)$ and $d(C_6, aba)$) to the respective reference atom and $\mu = 2.0$. The highlighted values represent the best possible alignment of the *aba-bonds* for the atoms C_1 and C_6 determined by Equations (4.6) to (4.8).

aba-bond for C_1	*aba-bond* for C_6	$C_6 - C_4$	$C_4 - C_2$	$C_4 - N_5$	$C_2 - C_1$	$C_2 = C_3$
	$d(C_6, aba)$ \ $d(C_1, aba)$	0	1	1	2	2
$C_1 - C_2$	0	**1.000**	0.111	0.011	0.040	0.032
$C_2 = C_3$	1	0.089	0.200	0.020	0.050	**0.063**
$C_2 - C_4$	1	0.111	**0.250**	0.025	0.063	0.050
$C_4 - N_5$	2	0.004	0.006	**0.063**	0.011	0.009
$C_4 - C_6$	2	0.040	0.063	0.006	**0.111**	0.089

each *aba-bond* of each atom must be matched to the best possible *aba-bond* of the other atom. Applying the Kuhn-Munkres algorithm allows one to discover which *aba-bond* (if any) of C_1 best matches each *aba-bond* of C_6. On the same table, the highlighted values indicate the best possible assignments. The sum of the best possible alignment score between *aba-bonds* is the actual *aba-bond* similarity score (Equations (4.3) to (4.5)) between these 2 atoms, thus: $T(C_1^A, C_6^A) = 1.000 + 0.250 + 0.063 + 0.111 + 0.063 = 1.487$.

To compute the self similarity score of molecule A, which represents the self-superimposition of the molecule, the procedure would be repeated for each atom and to obtain the final score for the atom matching use equations (4.6) to (4.8). Yet to better illustrate the algorithm as a tool for inferring structural similarity between molecules, this procedure will be demonstrated for comparing two different molecules A and B (Figure 4.3). However, it must be stressed that molecule

Table 4.5: Atom similarity scores between molecules A and B calculated by Equation (4.3). The highlighted values represent the best possible alignment of the atoms of molecule A and B determined by Equations (4.3) to (4.5).

Mol. A \ Mol. B	N_1	C_2	C_3	C_4	C_5
C_1	0.49	1.44	1.38	1.35	**1.39**
C_2	0.41	1.40	**3.05**	2.00	2.00
C_3	0.46	1.20	1.21	**1.39**	1.19
C_4	0.43	**2.40**	2.20	1.89	2.09
N_5	**1.43**	0.46	0.43	0.57	0.53
C_6	0.53	1.36	1.51	1.19	1.19

B includes a cyclic element, and therefore another characteristic was also considered in the W_m function beyond those already defined in the rules (4.10) to (4.12):

$$W_4 = \begin{cases} 0.8 & \text{if one atom is within a cycle and the other one is not} \\ 1.0 & \text{if both atoms are within a cycle} \end{cases} \tag{4.13}$$

Adding this new rule (4.13) to the others (4.10) to (4.12), the end result of the atom matching matrix was calculated and is presented in Table 4.5. Similarly to the bond matching process, the Kuhn-Munkres algorithm was used to match the atoms of both molecules and produce an optimal atom-matching score to the system defined by Equations (4.6) to (4.8).

As graphically shown in Figure 4.4, the atom alignment has preserved the obvious characteristics of the molecules while adequately handling the cyclic element. Also, atom C_6 of molecule A was not paired, as no adequate matching for this atom was found in molecule B.

The total molecule similarity score is then $S(A, B) = 1.43+2.40+3.05+1.39+1.39 = 9.66$. Using the same process, the self-similarity of molecules A and B was calculated as $S(A, A) = 13.89$ and $S(B, B) = 12.69$, respectively. These values allow us to use the Equation (4.9) for attributing a final structural similarity coefficient between both molecules: $JC_{AB} = 9.66/(13.89 + 12.69 - 9.66) = 0.571$, or 57.1% similarity between both molecules. This value, even for this simplisti-

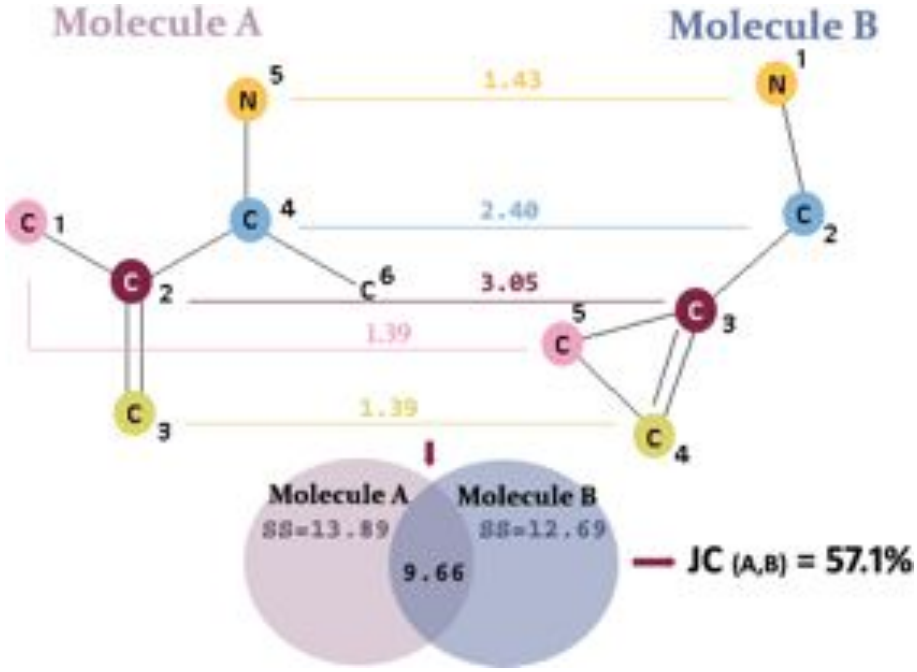


Figure 4.4: Atomic alignment between molecules A (*3-methylbut-3-en-2-amine*) and B (*cyclopropenyl methanamine*) and atomic and molecular similarity scores.

cally defined system, allows us to conclude that, even though an adequate atom matching between these two molecules was found, the structural similarity between both molecules is not very large, as expected by visual observation of their molecular graphs.

4.3 Implementation

Usually, if you have a new idea, you very rarely break through to anything like recognizable development or implementation of that idea the first time around - it takes two or three goes for the research community to return to the topic.

∼ Martin Fleischmann, Infinite Energy (1996)

Binary fingerprints, computed from the presence or absence of molecular features are commonly compared using a similarity coefficient as a measure of similarity between structures, with the Tanimoto coefficient being the most widely used (Flower, 1998). This is a particularly efficient, simple and among the most widely used methods in the case of two-dimensional or other easy to calculate

descriptors due to their performance (Flower, 1998; Heikamp & Bajorath, 2011; Willett, 2004). In addition, binary representations are suited to computer processing with a fast paced process. Binary fingerprints and similarity quantification using the Tanimoto coefficient have some drawbacks and limitations (Bender & Glen, 2004; Flower, 1998; Willett *et al.*, 1998) such as (1) not taking into account bits that are off in both molecules; (2) do not consider the frequency of detected fragments, therefore binary fingerprints tend to favour larger molecules in similarity due to bit saturation and smaller molecules in diversity selections; (3) in the cases where the bond path length exceeds the defined maximum, it is not possible to discriminate such fragments; (4) in the case of hashed fingerprints it is possible that different fragments hash to the same bit and therefore there is information loss; and (5) not including stereochemistry information which considerably tends to influence the properties/activities of the molecule. Nevertheless, there are several studies in the literature showing that fingerprint-based methods outperform other types of descriptors and similarity methods (Brown & Martin, 1996, 1997; Delaney, 1996; Martin *et al.*, 2002; Matter, 1997). We can hence implement NAMS in order to experimentally evaluate its capacity to compare molecules by examining and comparing the results for three different case-studies with another implementation based on the widely used daylight fingerprints.

4.3.1 Implementation of Fingerprints-based Structural Similarity

The implementation of daylight fingerprints is detailed in Appendix A - Descriptor set **B** (A.2.4). These fingerprints encode the substructures present in a molecule, which can then be compared in order to obtain the proportion of substructures in common between the two molecules under consideration. As explained above, in this study the Tanimoto coefficient was used, since it is the most widely used similarity coefficient for comparing binary fingerprints by establishing a ratio between the number of chemical features that are common to both molecules compared to the number of chemical features that are present in both molecules.

4.3.2 Implementation of Non-contiguous Atom Matching Structural Similarity

NAMS was implemented in Python (version 2.7). To process the chemical structures and extract the set of attributes to characterize molecule's atoms, Open-Babel libraries (version 2.3.1) were used (O'Boyle *et al.*, 2011). Although several code optimizations were used, this implementation was essentially designed for functionality and tool integration and not for computational performance.

The current implementation uses seven distinct atom and bond characteristic parameters, necessary for the W_n functions of eq. (4.1). These parameters are:

- *the nature of the atomic elements*: to compare and score atoms based on their type several *atom substitution matrices* (ASM) were devised. These matrices are inspired on the BLOSUM (Henikoff & Henikoff, 1992) substitution matrix that have been previously used to score alignments between evolutionarily divergent protein sequences, however in the context of this work they represent the distance (normalized between 0 and 1, where 0 represents atoms 100 % similar and 1 represents atoms completely different) between the elements (represented by their atomic numbers). Specifically five different matrices were devised (Appendix B.8): (1) ASM = 0, each atom type is only fully similar to itself (distance = 0) and completely different from all the others (distance = 1); ASM = 1, each atom type is only fully similar to itself (distance = 0) and partially different from all the others (distance = 0.9); ASM = 2 and 3, each atom type is only fully similar to itself (distance = 0) and partially different from all the others according to their position in terms of group and period in the periodic table (e.g. halogens are more similar); ASM = 4, all atoms are 100 % similar (distance = 0). We defined different matrices because there is not a standard way to compare atoms and furthermore, the nature of the study requires different needs (e.g. studies focused on the topology of the molecule should not account for the atom label (ASM = 4), while studies for which the properties of the atoms are key should not only take into account a binary comparison of the labels (equal or not equal as for ASM = 0 or 1) but also the existence of atoms with similar properties (ASM = 2 or 3));

- *whether the atom is chiral and its orientation*: a chiral tetrahedral carbon atom has four unique chemical groups arranged around it. The automatic detection and classification of chirality is a challenging task since it requires group theory (study the algebraic structures to distinguish stereoisomers using mathematical information about its symmetry), topology and geometry of a molecule (Gakh *et al.*, 2011). In order to establish the configuration, the four groups surrounding the central carbon atom (stereocenter) are ranked according to a priority sequence, determined by a number of sequence rules. The official systematic nomenclature to determine such rules was proposed in 1966 by Cahn, Ingold and Prelog (CIP rules) (Cahn *et al.*, 1966) and later extended (Prelog & Helmchen, 1982). The CIP rules play a double role: first, they allow determining whether the considered atom is really asymmetric, and second they rank the ligands connected to the stereocenter producing a pre-defined priority. We developed an algorithm that determines the chirality of a chemical structure based on widely used linear notations as input such as SMILES or InChI that was implemented in python (version 2.6) and uses OpenBabel-Pybel libraries (version 2.3.1) (O'Boyle *et al.*, 2011, 2008) for processing chemical structures. To determine the chirality according to the R-S notation several steps are needed, these can be summarized as follows: (1) indentify the stereocenters, (2) number the atoms in the molecule skeleton, (3) assign the priority of each ligand according to the CIP rules, (4) map the ligands into the skeleton groups which results in a permutation and finally (5) determine the parity of the permutation which allows the classification of the stereocenter(s) in R or S. The basic steps for describing a chiral center implemented in the algorithm are exemplified in Figure 4.5 for the compound (S)-1-amino-1-bromoethanol. The implementation of the algorithm and its exemplification are thoroughly described in a technical report (Teixeira *et al.*, 2013a);

- *the atom chain type*: allows the differentiations of atoms that belong to a linear, monocyclic or polycyclic chain;

- *the bond order*: is the number of bonding pairs of electrons between a pair of atoms. In a covalent bond between two atoms, a single bond has a bond

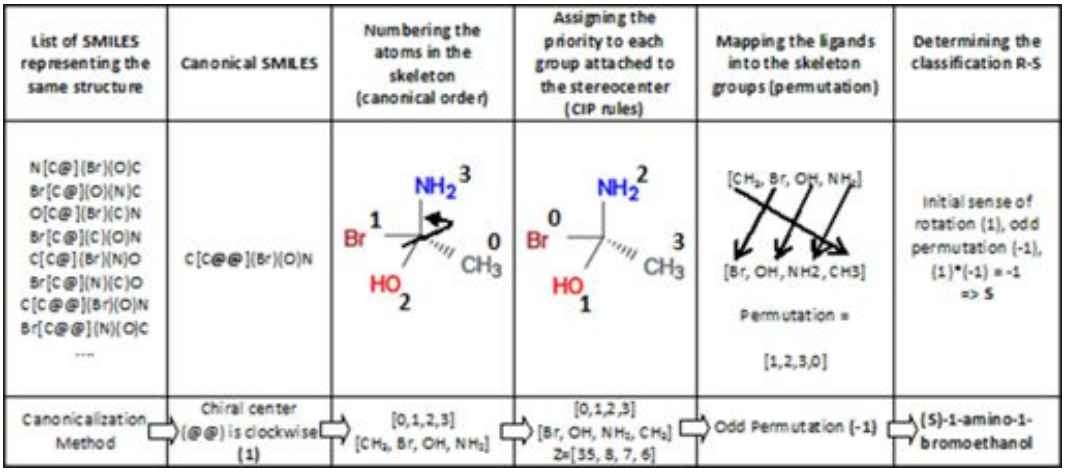


Figure 4.5: Basic steps for detecting and classifying a chirality center of the compound 1-amino-1-bromoethanol using the developed algorithm : the input notation is converted in a canonical SMILES, the initial sense of rotation is determined, at the stereocenter the atoms are then separated into the skeleton and its ligands and skeleton both are numbered independently in order to determine the permutation, finally the stereocenter can be classified as R or S combining the parity of the permutation and the initial sense of rotation.

order of one, a double bond has a bond order of two and a triple bond has a bond order of three;

- *the bond chain type*: allows the differentiations of bonds that belong to a linear, monocyclic or polycyclic chain;

- *whether the bond is aromatic or the atom is within an aromatic structure*: aromaticity describes the way in which a conjugated ring of unsaturated bonds, lone pairs, or empty orbitals exhibits a stabilization stronger than what would be expected by the stabilization of conjugation alone. Aromaticity can also be considered a manifestation of cyclic delocalization and of resonance. This is considered to be because electrons are free to cycle around circular arrangements of atoms that are alternately single- and double-bonded to one another. These bonds may be seen as a hybrid of a single bond and a double bond, each bond in the ring is identical to every other.

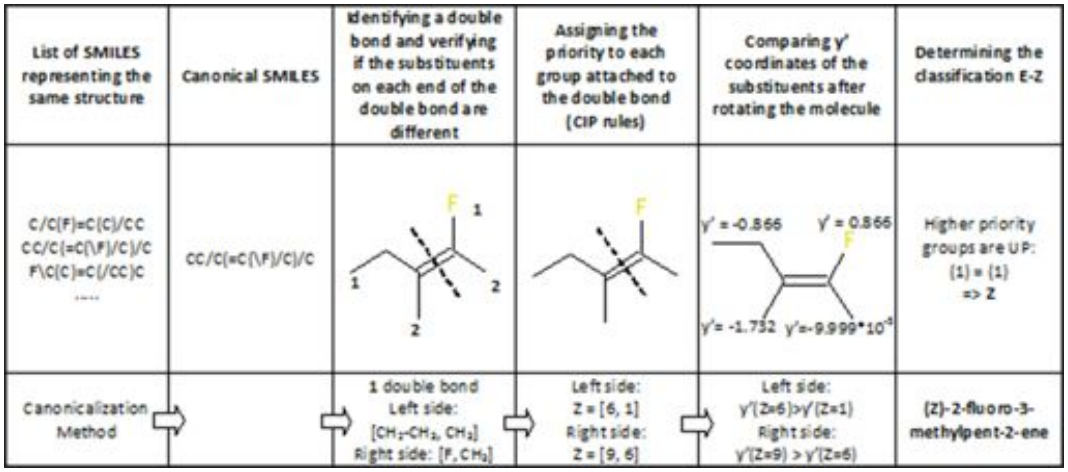


Figure 4.6: Basic steps for detecting and classifying E-Z isomerism in the compound 2-fluoro-3-methylpent-2-ene using the described algorithm: the input notation is converted in a canonical SMILES, the presence of a double bond and the substituents on each end of the double bond are verified, the priority of the substituents attached to each end of the double bond is determined according to the CIP rules, finally the double bond can be classified as E or Z comparing the y coordinates of the substituents of higher priority after rotating the molecule. In this case, the highest-priority groups on each side of the double bond are on the same side of the double bond. Fluorine is the highest priority group on the right side of the double bond, and ethyl is the highest-priority group on the left side of the molecule. This molecule can be classified as Z and the proper name is (Z)-2-fluoro-3-methylpent-2-ene.

- *whether a double bond has E-Z stereoisomerism and its orientation*: the geometrical (cis-trans) stereoisomerism arises when substituents are arranged differently in space due to restricted rotation of a double bond in a molecule (Mislow & Siegel, 1984). In a stereo bond, the substituents on either end of the double bond have to be different. To each substituent on a double bond is assigned a priority based on the CIP priority rules (Cahn *et al.*, 1966) and then classified according to the latest IUPAC recommendation (R Panico & Richer, 1993), the E-Z convention (Mislow & Siegel, 1984). To determine the double bond stereoisomerism according to the E/Z notation several steps are needed, these can be summarized as follows: (1) identify a double bond between two carbon atoms, (2) verify if the substituents on either end of the double bond are different, and (3) assign the priority of each

substituent on either end of the double bond according to the CIP rules. As final step in the procedure, compare the 2D coordinates of the substituents on either end of the double bond in order to determine if the substituents of higher priority are on the same or opposite sides of the double bond which allows the classification in Z or E, respectively. The basic steps for describing E-Z isomerism implemented in the algorithm are exemplified in Figure 4.6 for the compound (Z)-2-fluoro-3-methylpent-2-ene and detailed below. The implementation of the algorithm and its exemplification are thoroughly described in a technical report (Teixeira *et al.*, 2013a).

Table 4.6: The weights (W_m) of all characteristics under consideration for the similarity calculation for each case-study (**A1**, **E** and **F**). Atom and bond characteristics under consideration: (1) the nature of the element; (2) whether the atom is chiral and its orientation; (3) whether the atom is part of at least one ring; (4) the bond order; (5) whether the bond is part of at least one ring; (6)whether the bond is aromatic; (7) whether a double bond has E-Z stereoisomerism; (8) a penalty function to account for unmatched atoms.

Dataset	μ	Hydrogens	Atom and Bond Characteristics *							
			(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A1	2.0	No	†	0.95	0.7	0.7	0.7	0.7	0.95	-0.2
E	2.0	Yes	†, ‡	0.95	0.7	0.7	0.7	0.7	0.95	-0.2
F	2.0	No	‡	0.9	0.8	0.7	0.8	0.8	0.95	-0.2

†Atom substitution matrix that considers that each atom type is only fully similar to itself and completely different from all the others (Appendix B.8).
‡Atom substitution matrix with different scores computed according to their position on the periodic table (Appendix B.8).

Finally, two parameters not specific of *aba-bond* characteristics are also user defined: namely, the μ parameter (eq. (4.2)) and a penalty parameter that accounts for unmatched atoms, when the atom counts of molecules differ. The software also allows the user to specify whether hydrogen atoms should be accounted in the molecular comparison procedure. This has no effect in the use of the method, but it largely increases the computational cost.

The implementation of NAMS is available for download as a Python module or raw source code at `https://pypi.python.org/pypi/NAMS/0.9.2` along with several usage examples.

Table 4.6 displays the parameters used for the similarity calculation for each case-study (A1, E and F) analysed (Appendix chapter:cs).

4.3.2.1 Computational efficiency

Being developed in a scripting language, the current prototype is not very efficient computationally. Yet, several optimizations were performed to allow its use in moderately sized databases. One of the most important optimizations involved identifying the bottlenecks of the algorithm. This was clearly the *aba-bond* matching procedure (equations (4.3) - (4.5)) that used the Kuhn-Munkres algorithm. The $O(n^3)$ complexity can be a significant factor when comparing large molecules, and therefore a sub-optimal strategy was deployed that involved the use of a fast simple greedy heuristic for *aba-bond* matching. This heuristic, in general produces results similar to the Kuhn-Munkres algorithm, yet is up to four times faster on average. The Kuhn-Munkres algorithm returns an optimal solution and presents the best compromise between correctness and execution time, however the results produced by tests in these three case-studies using the greedy heuristic are only on occasion different from the ones reached using Kuhn-Munkres, and have never accounted for differences above 3 % in the final similarity score between 2 molecules.

A systematic test of NAMS over several databases in a common desktop PC (CPU Intel Core i3, running at 3.0 GHz with 4 GB of RAM) produced average computation times of 210 ms for comparing 2 molecules when using the Kuhn-Munkres algorithm. The heuristic processing times were on average 55 ms per pair of molecules compared. The heuristic approach was globally used for all the datasets in this study, however the Kuhn-Munkres was always used for the atom matching procedure (equations (4.6) - (4.8)).

4.4 Results and Discussion

All things are the same except for the differences, and different except for the similarities.
∼ Thomas Sowel, The Vision of the Anointed (1996)

Assessing a similarity function is a challenging task, since similarity between two molecules is highly subjective, and even chemists are not consistent when comparing molecules (Lajiness *et al.*, 2004). For the assessment of NAMS against the molecular similarity calculated by comparing path-based fingerprints with the three datasets described above, three main points were considered important to evaluate: (1) identification of the most informative representation of molecular structures (avoid information loss); (2) fine granularity of the similarity score in order to be able to distinguish similar molecules and (3) verify the molecular similarity principle (Johnson & Maggiora, 1990) which states that structurally similar molecules tend to have similar properties (physical, chemical or biological) more often than structurally dissimilar ones.

4.4.1 Case-Study A1 - Discriminate molecules with repeated substructures

Hydrocarbon fragments are present in most types of compounds, consequently a good similarity method should be able to distinguish hydrocarbons with similar structures. Therefore, 100 compounds randomly selected from dataset **A1** (described in Appendix A - A.1.1.1) are used to evaluate the distribution of similarity scores between all pairs of molecules using fingerprints and NAMS. This dataset (Appendix B.1) presents an important challenge which is dealing with repeated substructures.

Figure 4.7 displays the distribution and variation of the pairwise similarity between the 100 hydrocarbons, totalizing 4950 different pairs of structures (excluding self-similarities) using fingerprints and NAMS (Appendix B.1). Figure 4.7 shows that using fingerprints to calculate the similarity between the pairs of molecules obtains a density curve with three distinct peaks, one with a mean value of 100% similarity which is the maximum value this distribution reaches, one with a mean value of 85.7% similarity which is slightly below the mean value of this distribution (89.1% $\pm$ 9.6%) and a smaller peak with a mean value of 71.4% similarity which is the minimum value of similarity obtained using fingerprints. On the other hand, using NAMS to calculate the similarity between the

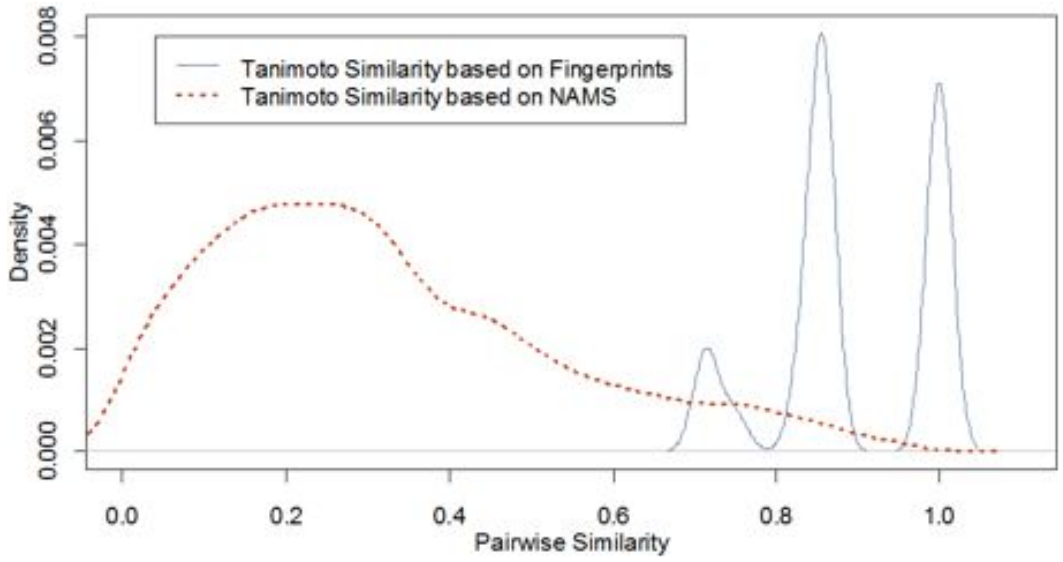


Figure 4.7: Distribution and variation of the pairwise similarity for a total of 4950 pairs (comparing the 100 hydrocarbons) using daylight fingerprints and NAMS.

pairs of molecules (Figure 4.7) obtains a continuous density curve. This distribution ranges from 0 to 97.2% similarity with a mean value of 31.8% ± 21.2% similarity. Comparing with fingerprints for which only 5 different values of similarity between the 4950 pairs where obtained (71.4%, 75.0%, 83.3%, 85.7% and 100%), using NAMS 859 different values of similarity are obtained showing its discriminative power. It is also interesting to mention that a similarity score between a pair of molecules of 100% was never obtained with NAMS, since one of its fundamental assumptions is that a molecule should only have a 100% similarity score when compared with itself.

Considering the question of effectiveness, i.e. being able to differentiate between molecules that are structurally different, Figure 4.8 gives an example of the similarity scores obtained using fingerprints and NAMS for four considerably structurally different compounds (Figure 4.8 I) and four considerably structurally similar compounds (Figure 4.8 II). The similarity score using fingerprints for all the molecules in the example is 100%, therefore there is no discriminative power between these structures. On the other hand, using NAMS to compare the considerably different structures (Figure 4.8 I), it is possible to verify that for example the structure *a)* is more similar to the structure *b)* than any of the remaining

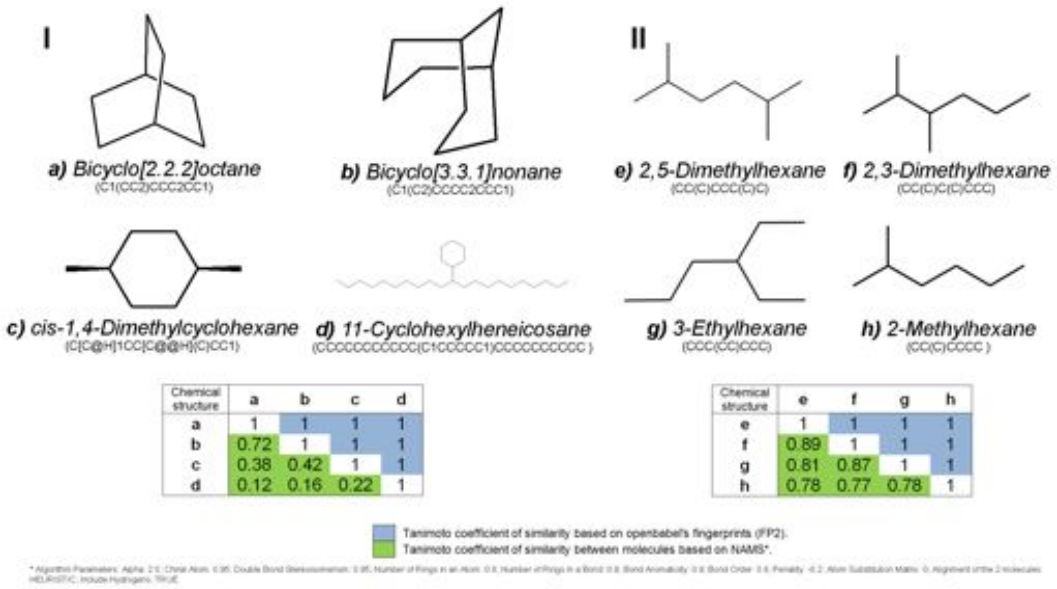


Figure 4.8: Example of the pairwise similarity using fingerprint and NAMS, obtained for 8 compounds extant in the dataset A1. The upper part of the pairwise similarity scores matrix (blue background) was calculated using fingerprints while the lower part of the pairwise similarity scores matrix (green background) was calculated using NAMS. **(I)** Four considerably structurally different compounds: *a)* Bicyclo[2.2.2]octane (SMILES C1(CC2)CCC2CC1); *b)* Bicyclo[3.3.1]nonane (SMILES C1(C2)CCCC2CCC1); *c)* cis-1,4-Dimethylcyclohexane (SMILES C[C@H]1CC[C@@H](C)CC1) and *d)* 11-Cyclohexylheneicosane (SMILES CCCCCCCCCC(C1CCCCC1)CCCCCCCCCC). **(II)** Four considerably structurally similar compounds: *e)* 2,5-Dimethylhexane (SMILES CC(C)CCC(C)C); *f)* 2,3-Dimethylhexane (SMILES CC(C)C(C)CCC); *g)* 3-Ethylhexane (SMILES CCC(CC)CCC) and *h)* 2-Methylhexane (SMILES CC(C)CCCC).

structures, since both have two fused rings and similar size. Using NAMS to compare the similar structures (Figure 4.8 II) it is possible to distinguish them in terms of shape and size. The structure *h)* has 7 carbon atoms, while all the others have 8 carbon atoms, therefore it is the structure with lowest similarity scores when compared with the others. The structures *e)* and *f)* are the most similar ones, since the only difference between them is the position of one methyl group. The structure *g)* is more similar to the structure *f)* because both have a substituent group in the position 3, a methyl in *f)* and an ethyl in *g)*.

4.4.2 Case-study E - Discriminate similar molecules with different activity levels

The case-study **E** (described in Appendix A - A.1.5) was chosen because although the structures have a similar structure, their activity level ranges from -5 to -7.881 with a mean value of -6.384 ± 1.082 (Appendix B.5). Considering that the binding strength of a receptor-substrate complex strongly depends on the shape of the substrate, the aim is to analyse the difference in the binding affinity of each pair of steroids to the corticosteroid binding globulin (CBG) receptor solely based on their similarity.

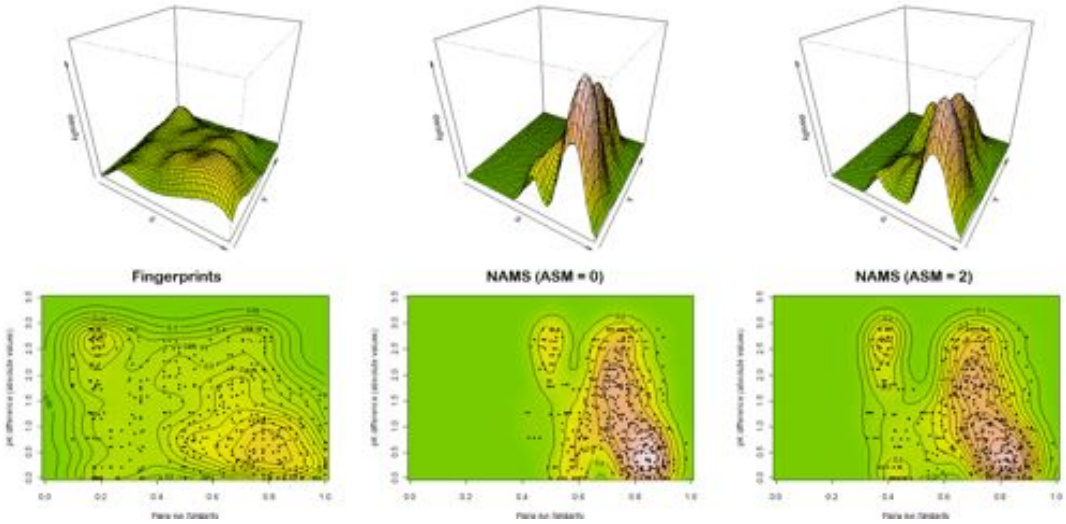


Figure 4.9: 2D Kernel density estimator perspective and contour plots showing the distribution of the pairwise similarity between the 31 steroids, totalizing 465 different pairs of structures (excluding self-similarities), calculated using fingerprints and NAMS (using two different atom substitution matrices (Appendix B.8 - ASM=0 and 2) and the corresponding difference in the pK absolute value.

Figure 4.9 displays the 2D kernel density estimator perspective and contour plots showing the distribution of the pairwise similarity between the 31 steroids, totalizing 465 different pairs of structures (excluding self-similarities), calculated using fingerprints and NAMS (with two different atom substitution matrices: the first one considers that each atom type is only fully similar to itself and completely different from all the others and the second one which considers different similar-

ities between the atoms, based on their position in the periodic table (Appendix B.8 - ASM=0 and 2) and the corresponding difference in the pK absolute value. While the pairwise similarity using fingerprints has a wider distribution, using NAMS concentrates the similarity between 35 and 98%. Using NAMS, there are two main zones with an high density of pairs of steroids, one between 70% and 80% of similarity and another one with higher similarity values (>80%) and the difference in the absolute value of the binding affinity becomes smaller with the increase of similarity. There is an isolated island of pairs of steroids using NAMS with similarity values between 40 and 50% for the atom substitution matrix 0 or 35 and 45% for the atom substitution matrix 2 and high differences in the absolute value of the binding affinity.

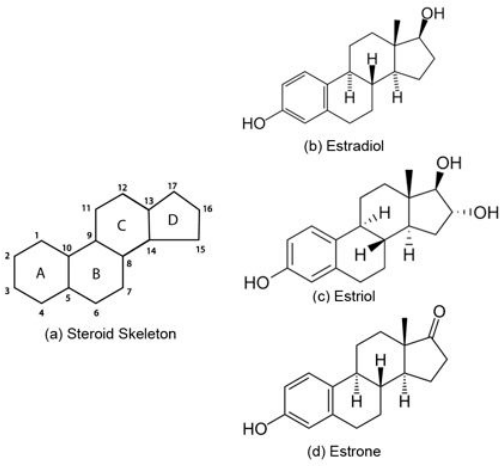


Figure 4.10: The (a) basic skeleton of a steroid and estrogenic steroids: (b) estradiol, (c) estriol and (d) estrone.

The pairs of compounds in this island were further investigated, leading to the conclusion that three of the compounds were always present in these pairs, namely estradiol, estriol and estrone (Figure 4.10). All these three compounds are estrogenic steroids, and although they share strong resemblance with the remaining steroids, there are some differences due to the aromatization process to convert anabolic steroids in estrogens. The "A" ring in the skeleton of a steroid (Figure 4.10 - a)) is now aromatic and it is a key functional group in all estrogens.

In general, it is possible to verify that although the structure of the 31 steroids is very similar, NAMS is able to discriminate them according to their pairwise similarity *versus* their difference in the binding activity level, since the density of points in Figure 4.10 is higher for higher similarity values *versus* lower difference in the binding activity level. NAMS was also able to discriminate the group of estrogenic steroids from the rest of the steroids. Using the fingerprints it is not possible to discriminate or relate the similarity between the molecules *versus* their difference in the binding activity level, since the distribution is wider and does not demonstrate any patterns.

4.4.3 Case-study F - Molecular similarity for inference

For dataset **F** (described in Appendix A - A.1.6), the aim is to retrieve compounds with similar activity level based on the similarity threshold which represents a multi-class classification task. For that purpose, using each compound as seed in each level of activity, the fraction of actives that are retrieved using Fingerprints and NAMS to measure the similarity between the compounds with different threshold cut-offs are recorded.

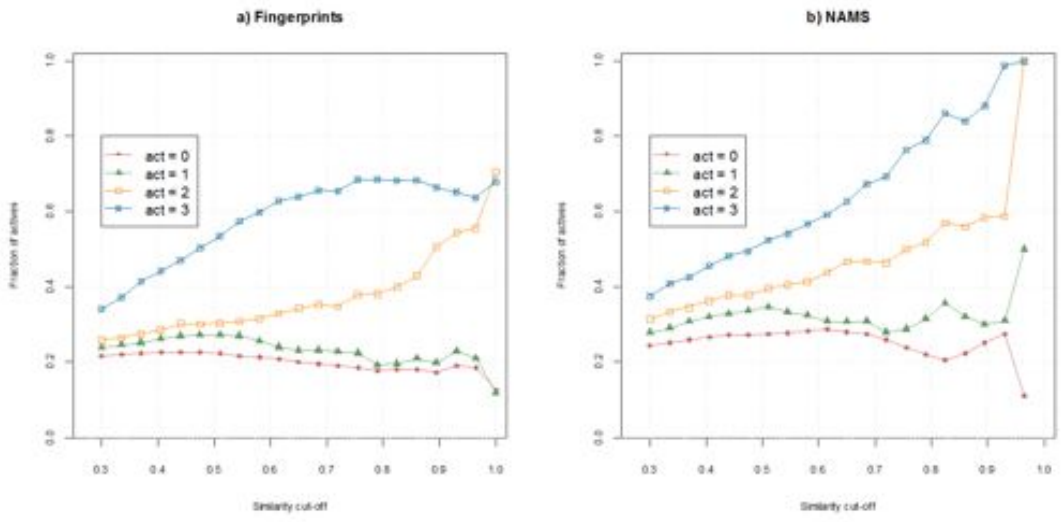


Figure 4.11: Fraction of active compounds in the dataset that are similar to seeds with a certain level of activity (act = 0, act = 1, act = 2, act = 3) used for similarity search with different threshold cut-offs (starting from 30%).

Figure 4.11 shows the fraction of actives within those compounds similar to compounds of each level of activity (0, 1, 2 or 3), given the minimum threshold of similarity for the search using Fingerprints (Figure 4.11 - a)) or NAMS (Figure 4.11 - b)). Figure 4.12 shows the fraction of actives (low, moderate or high activity) within those compounds similar to compounds that are active or inactive using Fingerprints and NAMS, given the minimum threshold of similarity for the search.

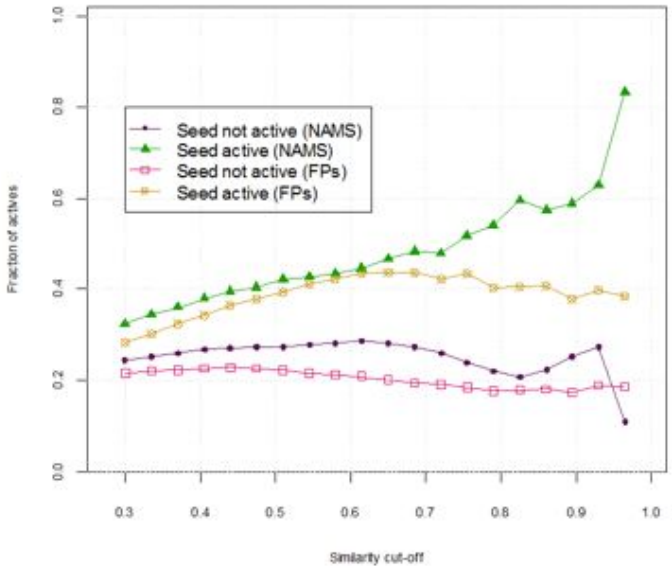


Figure 4.12: Fraction of actives compounds (activity level = 1, 2 or 3) within those compounds similar to compounds that are active (activity level = 1, 2 or 3) or inactive (activity level = 0) using Fingerprints (FPs) and NAMS with different threshold cut-offs (starting from 30%).

In general, using NAMS, the fraction of active compounds retrieved decreases as the similarity to an inactive is increased and increases as the similarity to an active is increased and the level of activity is higher. The same tendency is verified using Fingerprints, except when the seed has a low activity level (level 1), following the curve for inactive seeds. However, the fraction of actives retrieved when the seed is active is higher when using NAMS for the same similarity cut-off, particularly to higher levels of similarity. The fraction of actives within similar compounds of high activity (level 3) using NAMS is similar to using Fingerprints

until reaching a cut-off level of 70% of similarity, from this point on the fraction of actives retrieved is higher and increases to 1 at a cut-off level of 96.5% of similarity.

These results support the similarity principle, which states that compounds similar to biologically active ones should also be active and vice-versa, especially when the molecular comparison is highly discriminative.

4.4.4 Web Tool

A public and free Web tool for NAMS has been implemented. The objective was to produce an application simple to use with easily readable results that would allow the determination of the molecular similarity based on NAMS between (1) a pair of molecules, (2) a list of molecules or (3) a lead compound and a list of potential analogs. The user can input the molecules using their common name, SMILES string, or InChI identifier. The common name is resolved using the Chemical Identifier Resolver (NCI/CADD CIR, 2011), directly called by the application. The user can also define several parameters, according with the nature and objective of the problem, for the atom and bond structural characteristics, already described above, that will influence the atom/bond matching similarity score. It is important to note that to avoid ambiguity NAMS only considers double bond stereoisomerism or atomic chirality as characteristics if the stereo information is correctly and explicitly written in the molecule identifier. This webtool also allows the possibility to automatically analyse a molecule represented by its name, SMILES or InChI and generate a classification of the type of isomerism (chirality or double bond stereoisomerism) present in a given atom or bond (Teixeira *et al.*, 2013a).

The output produced consists of the molecule identifiers, a graphical representation of the molecular structure with the canonical numeration of atoms (generated using openbabel), the similarity score between each pair of molecules and a matrix of the atom similarity between both molecules, with the best possible alignment between the molecules highlighted (Figure 4.13). Furthermore, it is also possible to calculate the similarity score using different similarity/distance functions (e.g. Tanimoto, Cosine, Dice, Euclidean, etc) based on Fingerprints.

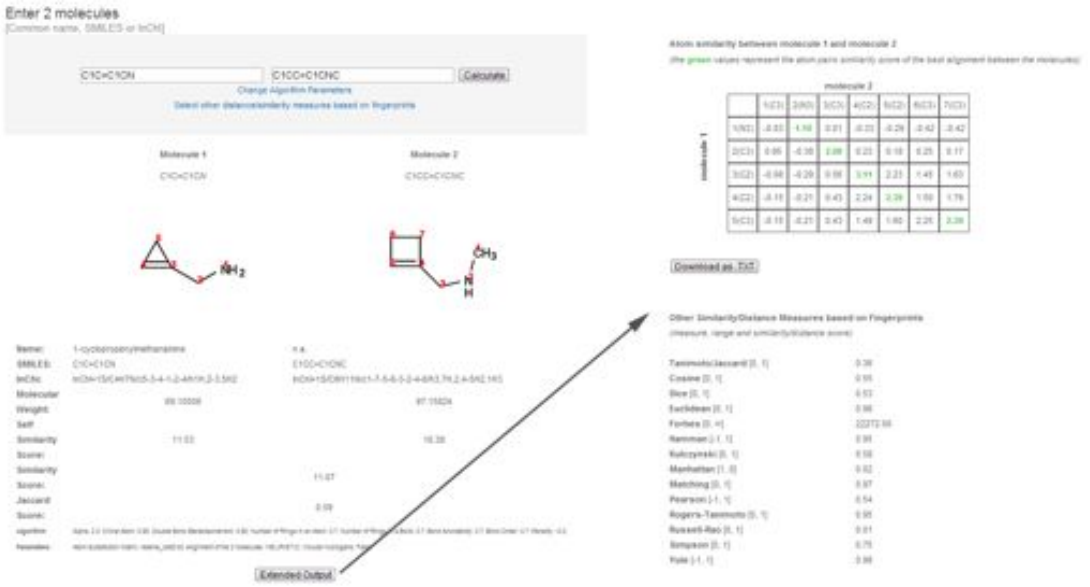


Figure 4.13: Screenshot of the NAMS Web-tool output when comparing two simple molecules: graphical representation of the molecules under comparison and molecule identifiers, self similarity scores for each molecule, similarity between both molecules, atom similarity between both molecules with the best possible alignment highlighted in green and finally other similarity scores using different similarity/distance coefficients based on Fingerprints.

This Web tool was developed, mainly, in the PHP programming language. The application communicates with the Python code that uses openbabel (O'Boyle *et al.*, 2011) for converting the different representations of the molecule and determining the similarity score of the molecules in accordance with the described algorithm. This Web tool also includes a section to download the raw source code or a python module that implements NAMS and is freely available at `http://nams.lasige.di.fc.ul.pt/`.

4.5 Summary

In this chapter, a new non-contiguous atomic alignment method for the analysis of structural similarity between molecules was defined and validated in three case-studies. The atomic alignment approach often requires high computational cost, however, the similarities detected by the atom correspondence are more intuitively

understood because similar atoms in the molecules are explicitly shown. This method is based on the comparison of atoms on comparable molecules taking into account their topological profiles and their intrinsic structural characteristics. The similarity measure is defined on the annotated molecular graph, based on a recursive concept of graph similarity and an optimal alignment between atoms using a heuristic and a penalty function to account for the differences in both atoms/bonds characteristics and topological profiles. The stereoisomerism and chirality are also considered in this similarity function since they are of great importance in many different fields since the molecular properties and biological effects of the stereoisomers are often significantly different. Considering that all similarity functions have a context that both define and limit their use, all defined atomic/bonds characteristics have a corresponding weight that can be adjusted or even eliminated in accordance with the context of the problem. New characteristics are also rather easy to include in the method.

The number of parameters used by NAMS may seem a deterrent for its use, but the tests made suggest that despite the fact that individual similarity scores do change, the similarity patterns are identical when comparing large databases. Also, the empirical tests over three case-studies presented strongly suggest that predefined default parameter values able to provide coherent results is attainable.

NAMS was compared with one of the most widely used similarity methods (Fingerprint-based similarity) for three case-studies with different objectives and characteristics. The method performed well and compared favourably to fingerprints for all 3 test cases. NAMS was able to distinguish either different or very similar hydrocarbons that were indistinguishable using a fingerprint-based approach and verifying the similarity principle using a dataset of very similar steroids with differences in the binding affinity to the corticosteroid binding globulin receptor. The method was also able to recover a significantly higher average fraction of active compounds when searching a database of highly diverse set of molecules with information about the MAO inhibition level. For this set it was verified that the fraction of actives recovered per active seed searched, consistently increased with the similarity level, which further suggests that NAMS is actually capturing reliable structure-activity relationships. Furthermore the main bottleneck in the application of this methodology is its computational cost which

is mainly due to the nature of the method. Future improvements in its execution time may be achieved by rewriting the current implementation of NAMS in the C programming language, and by using triangulation hierarchies (Jones & Ware, 1998) in order to implement neighbourhood search procedures.

It is nonetheless important to refer that although structurally-similar molecules are expected to exhibit similar properties, in some cases small changes in the structure of a molecule can bring thorough changes in some properties. Therefore, we cannot expect that in the context of property prediction there is a linear relationship between the molecular similarity of a pair of compounds and all the corresponding properties of that pair of molecules. A good similarity method is useful to construct a map of the chemical space, however this is not enough to make good property/activity predictions. The next chapter presents the development of tools able to analyse the chemical space defined by NAMS which may be able to recognize and make sense of structural patterns and their effects for property/activity prediction.

NAMS is made available freely for the whole community in a simple Web based tool at `http://nams.lasige.di.fc.ul.pt/`. The full source code of the Python module is also freely available within the same website.

Chapter 5

Instance-based Methods for Quantitative Structure-Property Relationship Modelling

The topic of this chapter is property prediction based on instance-based learning methods, which use instances to represent knowledge rather than pre-compiled general abstractions during prediction tasks. These algorithms are derived from the nearest neighbour pattern classifier (Aha & Kibler, 1991; Cover & Hart, 1967). The primary output of this type of methods is a function that maps structures to properties, i.e. given a structure drawn from the instance space it yields a prediction for its property value. Each prediction is based on the construction of a local approximation that is applied in the neighbourhood of the new target instance.

Classical model-based techniques in QSPR/QSAR problems have several shortcomings, namely (1) the predictive power of the model is highly dependent on the selection of predictor variables and on the presence of correlation between these variables, (2) the prediction capacity of the model is limited by the molecular diversity and distribution of the molecules in the training set (Oprea & Gottfries, 2001), (3) the models need to be re-trained every time new compounds are added or removed, and (4) usually only the uncertainty of the model is assessed and reported (Walker *et al.*, 2003).

5. INSTANCE-BASED METHODS FOR QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP MODELLING

Taking into consideration that structurally similar molecules tend to have similar properties (Johnson & Maggiora, 1990), we propose the use of a method that takes into account the high dimensionality of the chemical space, predicting chemical, physical or biological properties using Kriging (Isaaks & Srivastava, 1989) based on the most similar compounds in the molecular space composed by the instances of the training set and constructed based on their molecular similarity, consequently avoiding the selection of descriptors. Furthermore, the method takes into account the fact that similar molecules in the chemical space tend to yield similar property values while distant molecules can have very different values. However, the definition of the region of the chemical space that is adequately represented by similar compounds in the training set is not trivial, since the concept of similarity is subjective and even chemists are not consistent when comparing molecules (Lajiness *et al.*, 2004). The definition of similarity for molecules consists of mapping the chemical space, specifically representing the molecules and quantifying the similarity between them, enabling, in light of the similarity principle (Johnson & Maggiora, 1990), the use of the derived similarity measures in the prediction context. Various methods to define structural similarity between molecules are available in the literature (Maldonado *et al.*, 2006; Nikolova & Jaworska, 2003). This methodology uses ordinary kriging coupled with different molecular similarity approaches (based on molecular descriptors, fingerprints and atom matching) which creates an interpolation map over the molecular space that is capable of predicting properties/activities for diverse chemical datasets.

This chapter describes the steps to establish a quantitative structure-property relationship modelling using instance-based methods, namely (1) the use of structural similarity function to compute similarity between training molecules and test molecules; (2) the use of different similarity functions coupled with a Kriging algorithm to predict properties of chemical compounds for three case-studies of diverse chemical compounds collected from the literature; (3) the definition of a chemical neighbourhood in the prediction context and finally (4) a brief comparison between instance- and model-based learning methodologies.

5.1 From Similarity to Property Prediction

Classical QSPR/QSAR approaches, presented earlier in this document, have several shortcomings, namely (1) the predictive power of the model is highly dependent on the selection of predictor variables and on the presence of correlation between these variables, (2) the prediction capacity of the model is limited by the molecular diversity and distribution of the molecules in the training set (Oprea & Gottfries, 2001), (3) the models need to be re-trained every time new compounds are added or removed, and (4) usually only the uncertainty of the model is assessed and reported (Walker *et al.*, 2003).

For this study we propose the use of a method that, in light of the structural similarity principle (Johnson & Maggiora, 1990), takes into account the high dimensionality of the chemical space, predicting chemical, physical or biological properties based on the most structurally similar compounds in the molecular space, consequently avoiding the selection of descriptors. Another aspect we will address is the assessment of the reliability and the uncertainty of each estimation based on the structural similarity level. However, as already analysed in this document, the definition of structural similarity is not trivial, since the concept of similarity is subjective and even chemists are not consistent when comparing molecules (Lajiness *et al.*, 2004). The definition of structural similarity for molecules consists of mapping the chemical space, specifically representing the molecules and quantifying the similarity between them, enabling, in light of the similarity principle, the use of the derived similarity measures in the prediction context. Various methods to define structural similarity between molecules are available in the literature (Maldonado *et al.*, 2006; Nikolova & Jaworska, 2003) and they can be divided in three broad categories, each with its own specificities - approaches based on: (1) structural descriptors (two- and three-dimensional)(Basak *et al.*, 2003; Gute & Basak, 2001; Li & Colosi, 2012; Patterson *et al.*, 1996), (2) molecular fragments (such as fingerprints) (Flower, 1998) and (3) graph matching/descriptor-independent methods (such as the non-contiguous atom matching function (NAMS) (Teixeira & Falcao, 2013)).

Making predictions out of similarity or distance metrics is a known problem in several areas of science. One of the most known and used methods for quantita-

tive estimation based on topological distances is kriging, a method traditionally used in geostatistics which makes use of Tobler's first law of geography (Tobler, 1970): *"Everything is related to everything else, but near things are more related than distant things"*, meaning that a spatial dependence in the data is considered contrary to traditional statistical methods which assume that all data are independent. This method involves the estimation of a regionalized variable at a particular unsampled location by the weighted combination of the values of the neighbouring locations (Davis, 2002; Isaaks & Srivastava, 1989; Matheron, 1965). The use of this method has several advantages, namely: (1) estimates the estimation error along with the estimate of the property for each compound and this estimation error is minimized, therefore it is expected to be zero at the locations where experiments are performed and to grow with distance from these; (2) easier to comprehend than a black box model; (3) makes use of the distance/similarity between the compounds and it is not dependent on the selection of molecular descriptors; (4) fast enough to apply to a large data set; (5) searches for the relationship among measured properties rather than approximate the modelled system by fitting the parameters of the selected basis functions.

Kriging models are not new in chemoinformatics. Pioneer work was developed by Burden (2001) which demonstrated the applications of kriging in QSAR modelling for three datasets. Fang *et al.* (2004) used this technique for predicting boiling points of hydrocarbons and showed that kriging models could significantly improve the performances of the models by other existing methods. Obrezanova *et al.* (2007) applied kriging for the prediction of absorption, distribution, metabolism and excretion properties. Hawe *et al.* (2010) used kriging to predict the basicities of pyridines. Sun *et al.* (2011) showed that kriging models were able to outperform other methods in the development of predictive models for skin absorption. However, in all of these studies there was always an explicit use of chemical descriptors arbitrarily chosen according to the nature of the problems. To the best of our knowledge, none of the above approaches combined structural similarity with kriging methods for property/activity prediction of chemical compounds. In this study, we intend to demonstrate the application of kriging for molecular property estimation coupled with different similarity metrics based solely on the structure of the compound.

When building instance-based property prediction models, the general intuition is that the predictive performance will improve as more data is used in the model. However, there are situations where using only a subset of the data has advantages such as reducing time needed to fit the model, avoiding numerical inaccuracies and improving the robustness of the model, especially when the data is non-uniformly distributed over the whole chemical space under consideration. A critical issue is the design of a neighbourhood search strategy that will maximize the predictive performance of the method within a reasonable amount of computation time.

The general objectives of this study are: (1) to demonstrate that structural similarity functions can be useful to define the chemical space that is used to accurately predict properties/activities for diverse chemical compounds as yet unmeasured or even not synthesized, (2) to assess the extent to which kriging can be used to predict unmeasured properties of chemical compounds that were selected randomly or based on temporal characteristics using solely the metric map defined by structural similarity, (3) determine the uncertainty of each estimation, (4) to determine the effect of the training set size on the predictive results of the method and (5) assess different neighbourhood search strategies in order to maximize the predictive performance. Further potential applications of this methodology are illustrated by using three different structural similarity approaches based on molecular descriptors, fragments and graph matching to predict aqueous solubility and inhibition activity using two datasets of compounds with different structural characteristics. Different methods to select neighbourhoods using case-study A1 (prediction of enthalpy of formation in gas phase using distance-based kriging) will be studied and it will be assessed whether it is advantageous in terms of predictive results and computation time. The specific objectives of studying different neighbourhood search strategies can be summarized as follows, demonstrate that (1) as separation distance between pairs of compounds increases, the number of pairs that contribute to the prediction will decrease dramatically; (2) using a local neighbourhood for real-time prediction is advantageous in terms of CPU time to solve the kriging system; (3) the chance of violating basic assumptions of kriging and facing numerical inaccuracies will

be higher if one works with the whole data set instead of a local nearby neighbourhood; and (4) more data does not necessarily imply better performance or robustness. To identify the N surrounding data points that should contribute to interpolation different search strategies will be applied, namely (1) random selection of the neighbourhood; (2) moving neighbourhood based on a fixed number of compounds that are closest to the location of the target compound or a distance threshold; (3) moving neighbourhood based on a fixed number of compounds coupled with a minimum distance threshold or deletion of neighbours with negative weights; (4) size of neighbourhood defined based on different criteria; (5) size of neighbourhood defined based on different criteria using distance-defined shells of compounds; and (6) stepwise sequential selection of compounds based on a criteria. A brief comparison between instance- and model-based learning methodologies will also be performed.

5.2 Methods

This section presents the modelling methodology which is based on the kriging algorithm that requires the use of the chemical space based on the distance between molecules, as well as different strategies to define the optimal neighbourhood for each prediction. Three different ways to represent molecules based solely on their structure are studied in order to determine if it is possible to use their structural distance to estimate properties and which is the best way to calculate it in order to maximize the predictive power and minimize the number of neighbour compounds needed. In order to ensure minimal bias in evaluating the results an internal and external validation procedure was followed and is described, both for model selection as well as for final model assessment.

5.2.1 Modelling Methodology

The estimation of property values for which their properties were not experimentally determined and based solely on the structural similarity between the molecules is not sufficient. It is necessary to take into account the irregularities

in the property values, i. e. if the response variable surface has some spatial correlation then it is possible to infer the response in the immediate environment. One method of incorporating these concepts in the estimation model is to use kriging. Kriging is a family of estimators generally used in geostatistics for the interpolation of spatial data, i.e. to estimate variables at unobserved locations based on observed points at nearby locations (Isaaks & Srivastava, 1989; Matheron, 1965; Negreiros *et al.*, 2010). The kriging interpolation method seems to be a promising approach, as based on values measured in points from a certain range, it allows making predictions and the uncertainty of each prediction knowing just the distances to the known instances. The most widely used method is ordinary kriging, which was also selected for this study as it is the simplest model, makes no assumption on the nature or properties of the metric space and uses only distances between instances and measured values for inference.

5.2.1.1 Ordinary Kriging

The definition of Ordinary Kriging (OK) is often associated with the acronym *"BLUE"*, for *"best linear unbiased estimator"*. *"Best"* because OK aims at minimizing the variance of the errors, *"linear"* since its estimates are weighted linear combinations of the available data and *"unbiased"* because it attempts to reduce the mean residual error to zero. These goals are ambitious since the mean residual error and the variances of the errors are unknown for the data points to be predicted. When using other modelling techniques, the usual procedure involves building a model of the data and work with the average error and the error variance of the model. OK, on the other hand, uses a probabilistic model in which the bias and the error variance can be calculated in order to choose weights for the nearby sample which ensures that the average error of the model is zero and that the modelled variance is minimized. To estimate the error, its mean value and its variance, a random function model can be used, since it takes into account the uncertainty of what happens at unsampled points. This allows the construction of a map of both predicted values and level of uncertainty about the predicted values. To estimate unsampled points ($\hat{v}$) a weighted linear combination of the available samples can be used as in equation 5.1, where n is the number of compounds with

known property/activity in the set, v are the values of the property/activity, and w_j are the weights assigned to each known compound. The set of weights can change as the location of the unknown points change.

$$\hat{v} = \sum_{j=1}^{n} w_j \cdot v \tag{5.1}$$

The error of the i-th estimate (r_i) can then be defined as the difference between the estimated value ($\hat{v}_i$) and the true value at the same location (v_i) (equation 5.2).

$$r_i = \hat{v}_i - v_i \tag{5.2}$$

The average error (m_R) of a set of k estimates can then be defined as in equation 5.3.

$$m_R = \frac{1}{k} \cdot \sum_{i=1}^{k} r_i \tag{5.3}$$

However in practical situations the true value (v_i) is not known, therefore, as mentioned above, a probabilistic approach allows the calculation of unknown values as the outcome of a random process. For that purpose a random variable $V(x_0)$ with an expected value of E^2 is assigned to the unknown value to be estimated. The pairs of random variables have a distribution that depends only on their distance and not on their locations. The covariance between pairs of random variables separated by a distance h is $\tilde{C}_v(h)$. The predicted estimate and the estimation error are also random variables since these are the outcome of a weighted linear combination on the random variables at the available sample location as described in equation 5.1 with $v_i = V(x_0)$ and equation 5.2 with $r_i = R(x_0)$. For an unbiased estimation it is important to take into account that $E\{R(x_0)\}$ should be equal to zero, which means that the sum of weights has to be equal to one.

The error variance σ_R^2 of a set of k estimates can be expressed as equation 5.4 and it represents the function that OK aims to minimize.

$$\sigma_R^2 = \frac{1}{k} \cdot \sum_{i=1}^{k} (r_i - m_R)^2 \tag{5.4}$$

The average error (m_R) is assumed to be zero and therefore it can be eliminated from the equation. As already mentioned the true values are not known, therefore the same random function models are needed to minimize the variance of the modeled error $R(x_0)$. For that purpose, the variance of the error can be expressed as the variance of a weighted linear combination of random variables (equation 5.5).

$$Var\{\sum_{i=1}^{n} w_i \cdot vi\} = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \cdot Cov\{v_i v_j\} \tag{5.5}$$

Minimizing the variance of the error requires setting the n partial derivatives, namely the weights $w_1, ..., w_n$ to zero. This produces a system of n simultaneous linear equations with n unknowns for the n sample locations, having in mind the unbiasedness condition that the sum of weights has to be equal to one. The solution for this $n+1$ system of equations is not straightforward, however it can be solved using Lagrange multipliers (equation 5.6).

$$\begin{aligned} \frac{\partial \sigma_R^2}{\partial w_i} = 0 \Rightarrow \sum_{j=1}^{n} w_j \tilde{C_{ij}} + \mu = \tilde{C_{i0}} \quad \forall \quad i = 1, ..., n \\ \frac{\partial \sigma_R^2}{\partial \mu} = 0 \Rightarrow \sum_{i=1}^{n} w_i = 1 \end{aligned} \tag{5.6}$$

This solution will provide the set of weights and the mean value (the Lagrange parameter μ) that minimizes the modeled error variance (equation 5.7) under the constraint that weights sum to one.

$$\tilde{\sigma_R^2} = \tilde{\sigma^2} + \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \tilde{C_{ij}} - 2 \cdot \sum_{i=1}^{n} w_i \tilde{C_{i0}} + 2 \cdot \mu (\sum_{i=1}^{n} w_i - 1) \tag{5.7}$$

The system of equations represented in equation 5.6 can be expressed in matrix notation which is usually known as the OK system (equation 5.8).

$$\begin{bmatrix} \tilde{C_{11}} & \cdots & \tilde{C_{1n}} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \tilde{C_{n1}} & \cdots & \tilde{C_{nn}} & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ \vdots \\ w_n \\ \mu \end{bmatrix} = \begin{bmatrix} \tilde{C_{01}} \\ \vdots \\ \tilde{C_{n0}} \\ 1 \end{bmatrix} \tag{5.8}$$

$$C \cdot w = D$$

$$w = C^{-1} \cdot D$$

The vector D provides a weighting scheme of the distances using the covariances between all the sample locations (denoting $i = 1, 2, ..., n$) and locations where an estimation is needed (denoting 0). The higher the covariance between a sample and the location being estimated, the more that sample contributes to the estimation. The matrix C describes the covariances between all the sample pairs, bestowing information about the distribution of the available sample data. Therefore, C matrix readjusts the sample weight according to their clustering. Alternatively to the covariance between a sample and the locations being estimated, a closely related measure can be used to give the same information - the semivariance.

In terms of the matrices defined in 5.8, the minimized error variance (equation 5.7) can be expressed as equation 5.9, which is usually referred to as the kriging estimated variance.

$$\tilde{\sigma_R^2} = \tilde{\sigma^2} - w \cdot D \tag{5.9}$$

The kriging estimated variance takes into account four important factors and the interactions between them: (1) the number of samples used to make the estimation, it is expected that estimates based on many samples will be more reliable than those based on just a few; (2) proximity of the samples, as the average distance increases, the estimate becomes less reliable; (3) the spatial arrangement (clustering) of the samples around the test compound; and (4) the nature of the problem in terms of spatial continuity, smoothness and well-behaved variables will have better estimates than very erratic variables.

In summary, to minimize the modelled error variance, it is necessary to choose the $(n + 1)^2$ covariances that will describe the spatial distribution of the random

function model. The set of weights that produce an unbiased estimate with a minimum error variance (equation 5.9) can be simply calculated using the system of equations 5.8. The choice of the covariance or semivariance model to describe the spatial continuity is then a pre-requisite to apply OK.

5.2.1.2 Semivariogram

A semivariogram describes how the spatial continuity changes with distance between all pairs of sample locations, quantifying the spatial correlation. In practice, OK is usually implemented using the semivariogram rather than the covariogram because it has better statistical properties (Bohling, 2005). The construction of a semivariogram consists of two parts: an empirical semivariogram and a model semivariogram that will extend the estimations to locations where there are possibly no sample locations by fitting a function to the empirical semivariogram. The empirical semivariogram is constructed by calculating the semivariance ($\gamma(h)$) of each point in the set with respect to each of the other points, using equation 5.10 in relation to the distance between these points.

$$\gamma(h) = \frac{1}{2N(h)} \sum_{N(h)} (z_i - z_j) \tag{5.10}$$

$N(h)$ is the set of all pairwise distances (h), and z_i and z_j are data values at spatial locations i and j separated by h. Using all pairs of compounds on the semivariogram may complicate its interpretation and therefore it is usual to apply a binning process i.e., average semivariance data by distance intervals. An appropriate parametric model is then typically fitted into the empirical semivariogram and utilized to calculate distance weights for interpolation. Identifying the optimal model may involve running and evaluating a large number of models. Usually the model includes three parameters: (1) the *"nugget"* which represents the semivariance at distance zero due to microscale variations or low accuracy of the measurement; (2) the *"range"* which represents the distance at which semivariance levels off, that is to say the spatially correlated portion of the data; (3) the *"sill"* which represents the semivariance at which the mentioned levelling takes place. After a suitable semivariogram model has been selected, the kriging process is able to define a continuous surface for the entire study area using

weights calculated with the semivariogram model, as well as values and locations of the measured points. It is also possible to adjust the distance or number of measured points that are used for making predictions for each unknown value.

5.2.1.3 Neighbourhood Selection Strategies

The aim of trying different neighbourhood selection strategies is to select the training set that will produce the model that most accurately approximates the function underlying the data. However, in practise we do not have an explicit description of the underlying function, which makes it impossible to directly optimize this objective. Instead, we optimize certain criteria in order to establish a general search strategy that improves the quality of the resulting model. In this section, different methods and combinations of methods to select molecules from the training set are presented.

(1) Random Selection of the Neighbourhood (RSN)

This method randomly selects a certain number N of compounds of the training set without taking into account any of their properties or similarity to the test compound. The performance of this method is used as a reference for the performance of the other methods.

(2) Moving neighbourhood (MN) based on a fixed number of compounds or distance threshold

This method is based on a fixed number N of compounds that are closest to the location of the target compound. In practice, when prediction is required at some test point, only those observations within a given distance of test compound are used in the prediction. This so-called moving neighbourhood, therefore, contains a specified subset of all the observations. It can be defined by requiring a minimum number of observations N or by spanning a given distance threshold from the test compound. The first advantage of such a strategy is that solution of the Kriging equations involves a matrix of reasonably pre-defined size and there are guarantees that there are not too few or too many compounds at a defined distance threshold. The second advantage is that pragmatic and sensible models can be fitted to the data. To determine the number N of nearby compounds,

several neighbourhood sizes must be tried with the objective of maximize the cross-validated RMSE.

(3) Moving neighbourhood (MN) based on a fixed number of compounds coupled with a minimum distance (MD) threshold or deletion of neighbours with negative weights (NW)

Negative weights must be avoided and they arise when data close to the location being estimated screen outlying data. Negative weights, when interpreted as probabilities for constructing a local conditional distribution, are nonphysical. Also, negative weights when applied to high data values may lead to negative and nonphysical estimates. Among the various solutions to the problem of negative kriging weights in this study the direct correction is applied (Yamamoto, 2000). After solving the ordinary kriging system, negative weights are corrected according to specific algorithms, such as Froidevaux (1993), Journel & Rao (1996), and Deutsch (1996) will be considered. Froidevaux's correction resets all negative weights to zero; hence, this procedure simply removes the data corresponding to negative weights and rescale the remaining positive weights, thus some information is lost. The correction proposed by Journel and Rao only removes the sample with the largest negative weight. After solving the kriging system, a positive constant equal to the modulus of the largest negative weight (if any) is added to all weights. Deutsch has proposed an approach that zeroes negative weights. Moreover, this procedure also removes samples that have (1) a weight less than the average absolute magnitude of negative weights; (2) a covariance between the location being estimated and the location of the sample that is less than the average covariance between the location being estimated and the locations receiving negative weights.

Moreover, to avoid redundancy a minimal distance between any pair of selected training data was also defined.

(4) Size of neighbourhood defined based on different criteria (NC)

This method defines the size of the neighbourhood based on the optimization of different criteria namely, minimize the kriging estimated variance, the interpolated variance or the condition number. The kriging estimated variance, as already explained, is used as a quality indicator of the estimator and it is used to assign confidence levels of each prediction. However, kriging estimated variance is

independent on the property values of each data point. An alternative measure of the reliability of ordinary kriging estimates is the interpolated variance (equation 5.11), which accounts for both data configuration (w_i, the OK weights) and data values ($z(x_i)$, the estimated property and $z(x_o)$, the properties of the training molecules) (Yamamoto, 2000).

$$s_o^2 = \sum_{i=1}^{n} w_i [z(x_i) - z(x_o)]^2 \tag{5.11}$$

The minimization of the condition number is also an important criterion to avoid ill-conditioned semivariance or correlation matrices (large condition number) and thus susceptible to numerical inaccuracies.

(5) Size of the neighbourhood defined based on different criteria using distance-defined shells of compounds (NC-SHELLS)

This methods uses the criteria previously defined in (4), however instead of using all compounds in the training set, it defines distance- or fixed number-based shells of compounds.

(6) Stepwise sequential selection of compounds based on a criteria (SC)

This method used the criteria previously described in (4) to gradually incorporate new data in the neighbourhood that maximizes the gain (decrease in the kriging estimated variance, interpolated variance and condition number).

5.2.2 Implementation of Ordinary Kriging

5.2.2.1 *CoordKrig* - Coordinate based kriging

The R package geoR (Diggle & Jr, 2007; Ribeiro Jr & Diggle, 2001) has an efficient implementation of OK. However, this package requires the coordinates of the data points instead of their distances, since this package was designed for geostatistical data analysis in which typical data inputs are the coordinates of data locations and the data values. For that purpose, multidimensional scaling (through the function isoMDS of R package MASS (Venables & Ripley, 2002)) was used to transform the distances between the molecules into (XY) coordinates. These coordinates were then jittered uniformly on the regions around points with very

similar coordinates using the function jitter2D of the geoR package.To fit a model to the semivariogram a spherical function was considered adequate given data distribution and the method variofit of the package geoR was used to estimate its parameters (sill and range) that give the smallest value of the summation and use them as initial values for the minimization of the loss function using cressie weights. The nugget was fixed at zero and the spherical function was used to model the semivariance. Preliminary tests with several data sets showed that this function provided consistently good results.

5.2.2.2 *DistKrig* - Distance based kriging

Alternatively, and since OK derives predicted values based on the distance between points in space and the variation between measurements as a function of distance, an OK algorithm was implemented in R using as input a distance matrix between the molecules and following all the steps presented in the modelling methodology. However, for every molecule that we aim to predict the property value, a neighbourhood will be demarcated by a pre-defined number ($neighs$) of molecules in the training set that are nearest to the test molecule. To fit a model to the semivariogram, linear regression was chosen since this function shown to be suitable to model the data and simple to automate the process, using the R package lm (Chambers, 1992) and defining that the regression line will pass through the origin (defined in geostatistics as the nugget), since it is assumed that if two molecules are 100% similar, then it is expected that they share the same property value.

5.2.3 Molecular Representation

The use of kriging for non-spatial problems requires working on a metric space, where the distances between all existing elements can be computed. In this context, a requirement to apply kriging is that the molecules need to be represented in a map based on their dissimilarity. As presented above, various methods to define structural similarity are available in the literature (Bender & Glen, 2004; Nikolova & Jaworska, 2003) and can be divided in three board categories; approaches based on structural descriptors, molecular fragments and graph match-

ing. In the conducted study these three approaches to quantify the structural similarity of molecules, which are posteriorly transformed into distances, will be explored:

5.2.3.1 A. Structural similarity based on molecular descriptors

Molecular descriptors can be computed from the molecular structure encoding in numerical form chemical information contained in the molecule. In this work a set of 1666 molecular descriptors (2 and 3-Dimensional) was generated for each dataset using e-DRAGON (Tetko *et al.*, 2005; VCCLAB, accessed in 2011), a free online version of DRAGON. The 3D atomic coordinates of the lower energy conformation for the provided molecules were calculated using CORINA (Sadowski *et al.*, 1994). A preprocessing step was carried out where all zero variance variables (i.e. all the observations are the same) were removed and standardization was applied to transform each descriptor values to have zero mean and unit variance according to equation (5.12) where x represents the values of a molecular descriptor, $\bar{x}$ the mean value for descriptor x and σ its standard deviation. Each descriptor positions each abstract molecular representation in the descriptor space and the molecular dissimilarity can be measured calculating the Euclidean distance between all dimensions of the chemical space.

$$x_{standardized} = \frac{x - \bar{x}}{\sigma} \tag{5.12}$$

5.2.3.2 B. Structural similarity based on molecular fragments

In this work the structural similarity score based on molecular fragments is obtained by comparing path-based fingerprints (FP2) calculated by openbabel (O'Boyle *et al.*, 2011) using the Tanimoto coefficient (Flower, 1998). The FP2 binary fingerprints are bit strings that encode the presence or absence of topological patterns up to 7 atoms in a molecule and map them onto a bit-string of length 1024 using a hash function (similar to the Daylight fingerprints).

The degree of similarity given by the Tanimoto coefficient ($s(x, y)$) was converted to a degree of dissimilarity ($d(x, y)$) applying a monotonically decreasing transformation using the natural logarithm (equation 5.13).

$$d(x, y) = -ln(s(x, y)) \tag{5.13}$$

5.2.3.3 C. Structural similarity based on graph matching

A molecule can also be represented, using graph theory, as a labeled graph whose vertices correspond to the atoms and edges correspond to the covalent bonds. The representation of molecules using graphs has some advantages, namely, graphs are intuitive when representing a molecule since they are close to our understanding of a molecule, it is a descriptor-independent approach and they have a solid mathematical background with different existing techniques to compare labeled graphs (Ehrlich & Rarey, 2011). In this study we will use the non-contiguous atom matching structural similarity method (NAMS) (Teixeira & Falcao, 2013; Teixeira *et al.*, 2013a) which, as presented in the previous chapter, has proven to be useful for comparing molecular structures. Again, the degree of similarity given by NAMS ($s(x, y)$) was converted to a degree of dissimilarity ($d(x, y)$) applying a monotonically decreasing transformation using the natural logarithm (equation 5.13).

5.2.4 Model Validation

The described approach requires a parametrization step in order to predict properties of new compounds by selecting the most similar compounds from the training set. For that purpose, a leave-one-out (LOO) cross validation approach was followed which comprises for n samples, the creation of n different learning sets and n different test sets by taking all the samples except one as learning set and the sample left out as test set. The goal of this cross-validation is not only to select the best parameters, but also to estimate the expected level of fit of the approach to new data that is not used in the training set and to statistically ensure that the approach is sound. The cross-validated correlation coefficient (q^2), the percentage of compounds for which the estimation error is between an acceptable interval and the root mean squared error (RMSE) are performed to determine the goodness of fit of the model.

Yet to adequately assess the validity of each model (Tropsha, 2010), an external validation set is used to predict the properties of a set of instances, never used in the model development so as to adequately evaluate how well the model generalizes in a real world scenario. To assess the external predictive ability of each model, three statistics are used, namely: the predictive correlation coefficient (Q^2), the percentage of compounds for which the estimation error is between an acceptable interval and the RMSE.

5.3 Data

For this study three case-studies of diverse chemical compounds presented in Appendix A are used: **(A1)** predicting enthalpy of formation of gas phase for ThermInfo's dataset, **(B)** predicting aqueous solubility and **(C)** predicting Dihydrofolate Reductase (DHFR) inhibition activity.

5.4 Results

In order to validate the hypothesis that it is possible to predict a property of interest based on structural similarity/dissimilarity between the molecules, as described above, the kriging algorithm was tested in case-studies A1, B and C (see Appendix A) with different parametrizations coupled with three distance matrices (based on molecular descriptors, fingerprints and NAMS). The results obtained for the best parametrizations are summarized below.

5.4.1 Case A1 - Predicting enthalpy of formation of gas phase

The general intuition is that using more data will always result in a better model. For this case-study the aim is to verify if this preposition is true and in case the result is negative, determine the best neighbourhood selection strategy. Therefore, the properties of the 364 compounds (Appendix B.1) in this case-study (Appendix A.1.1.1) were predicted using all the remaining 363 compounds in the training set while applying a leave-one-out cross-validation strategy. The results,

using NAMS to calculate the distance between compounds, have no predictive value. Therefore, the initial intuition that using more data will result in a better model did not verify. Additionally, to verify if there is margin to improve the predictive results using a smaller neighborhood, each compound in the training set was tested using a neighbourhood of N compounds ranging from 1 to 363 and the prediction of the property with smallest error was selected for each compound. In contrast with the previous results, these new results are highly predictive with a RMSE of 14.01, a q^2_{cv} of 99.46% using an average of 47 compounds as neighbourhood and with 92.64% of the compounds being predicted with an error lower than $\pm$ 10 kJ/mol which is in agreement with the experimental margin of error for this property. The problem then is related with the neighbourhood selection (as this selection was done based on the knowledge about the observed property) to predict the property of each compound, in order to maximize the predictive results.

A preliminary study showed that using NAMS to represent the compounds in the chemical space yields better results than fingerprints or molecular descriptors. Furthermore, the implementation that uses distance between points in space (*DistKrig*) also showed better predictive performance. Taking into account that for this case-study, the focus is to study different methods to select neighbourhoods, only the results for the best configurations previously determined in the preliminary study are shown in Table 5.1. Detailed results are available in Appendix C.4.

The best results were obtained using a moving neighbourhood (MN) based on the 20 closest compounds. These predictive results clearly demonstrate an improvement in relation to use all compounds in the training set or random subset selections. The combination of MN with other strategies, such as removing compounds with negative weights or selection of compounds based on a criterion did not improve the predictive results.

Table 5.2 shows a summary of the best models obtained using NAMS coupled with Kriging and with another instance based methodology (k-Nearest Neighbours), as compared, to model-based approaches using the methodologies presented earlier in this document. It is possible to observe that kriging performs

Table 5.1: Summary of the best results (leave-one-out cross validation) obtained for the training set (dataset A1) using different neighbourhood selection methods: (1) Random Selection of the Neighbourhood (RSN); (2) Moving neighbourhood (MN) based on a fixed number of compounds or distance threshold; (3) Moving neighbourhood (MN) based on a fixed number of compounds coupled with a minimum distance (MD) threshold or deletion of neighbours with negative weights (NW); (4) Size of neighbourhood defined based on different criteria (NC); (5) Size of the neighbourhood defined based on different criteria using distance-defined shells of compounds (NC-SHELLS); and (6) Stepwise sequential selection of compounds based on a criteria (SC).

Neigh. Selec.	**Neigh. Size**	**Negative Weights**	**Min Distance**	**Criteria**	**RMSE**	q^2_{cv}	**% error** ± 10
RSN	15	Yes	0	None	100.839	0.691	10.1
MN	20	Yes	0	None	**55.325**	**0.916**	**36.2**
	10	Yes	0.025	None	57.422	0.911	32.5
	10	froidevaux	0	None	68.427	0.872	24.0
	10	deutsch	0	None	69.995	0.866	23.4
	10	journel	0	None	76.668	0.839	16.6
NC	364	deutsch	0	cond. num.	74.429	0.848	28.2
	364	deutsch	0	evar	85.848	0.798	25.6
	364	Yes	0	interp. var	75.253	0.845	34.7
NC-SHELLS	10	deutsch	0	cond. num.	71.201	0.861	27.8
	10	deutsch	0	evar	71.157	0.861	27.6
	10	froidevaux	0	interp. var	71.515	0.860	28.2
SC	10	Yes	0	cond. num.	69.472	0.868	33.4
	20	Yes	0	evar	61.176	0.898	37.5
	20	Yes	0	interp. var	82.416	0.814	31.7

Table 5.2: Comparison of the predictive power of the model developed in this study with other model-based approaches using different machine learning techniques (Random Forests, Support Vector Machines and Variable Importance as calculated by RFs for feature selection) and different combinations of descriptors for case-study A1.

Molecular Representation	Method	Val. Type	Neighborhood size	RMSE	q^2
NAMS	Kriging	LOO	20	55.325	0.916
NAMS	k-NN	LOO	5	80.45345	0.822832
A + B + D	RF-VI	10-f CV	89	34.1	0.9686
NAMS	RF-VI	LOO	234	46.80374	0.940351
NAMS	RF	OOB	364	70.45	86.42
NAMS	SVM	LOO	364	49.33807	0.933785

better than k-NN, however, most methodologies based on models show advantages in terms of predictive performance. Nevertheless, Kriging is able to obtain results within the same order of magnitude without requiring any feature selection step and allowing a clear interpretation of the results, as well as feedback on the kriging estimated variance for each prediction.

5.4.2 Case B - Predicting Aqueous Solubility

For the aqueous solubility dataset the distribution of the pairwise distance between the 1033 compounds in the training set was preliminarily analysed. Figure 5.1 plots the 533028 different pairs of structures (excluding self-distances) calculated using a) molecular descriptors, b) fingerprints and c) NAMS and the corresponding difference in the aqueous solubility absolute value.

In general, it is possible to verify that NAMS and fingerprints are able to discriminate the compounds according to their pairwise distance versus their difference in the aqueous solubility value and verify the similarity principle (Johnson & Maggiora, 1990), since the plot of pairwise distance values versus absolute difference in the aqueous solubility values (Figure 5.1 - c) and e)) for the set of molecules exhibit a trapezoidal distribution, revealing a neighbourhood behaviour with a low frequency of pairs in the upper left triangle (very similar compounds with a high degree of difference in the property value). In the probability distribution plot, it can be observed (Figure 5.1 - d) and f)) that for both Fingerprints and especially for NAMS there is a high probability for compounds that are very close to each other to have a small difference in the property value. While using molecular descriptors the relationship between the pairwise distance of the compounds and their difference in the property is not as clear and the discrimination of the compounds is more complicated, since there is a high density of pairs with high similarity values and high differences in the property value. The only tendency that is shown for molecular descriptors is that for high distance scores the property value is also dissimilar.

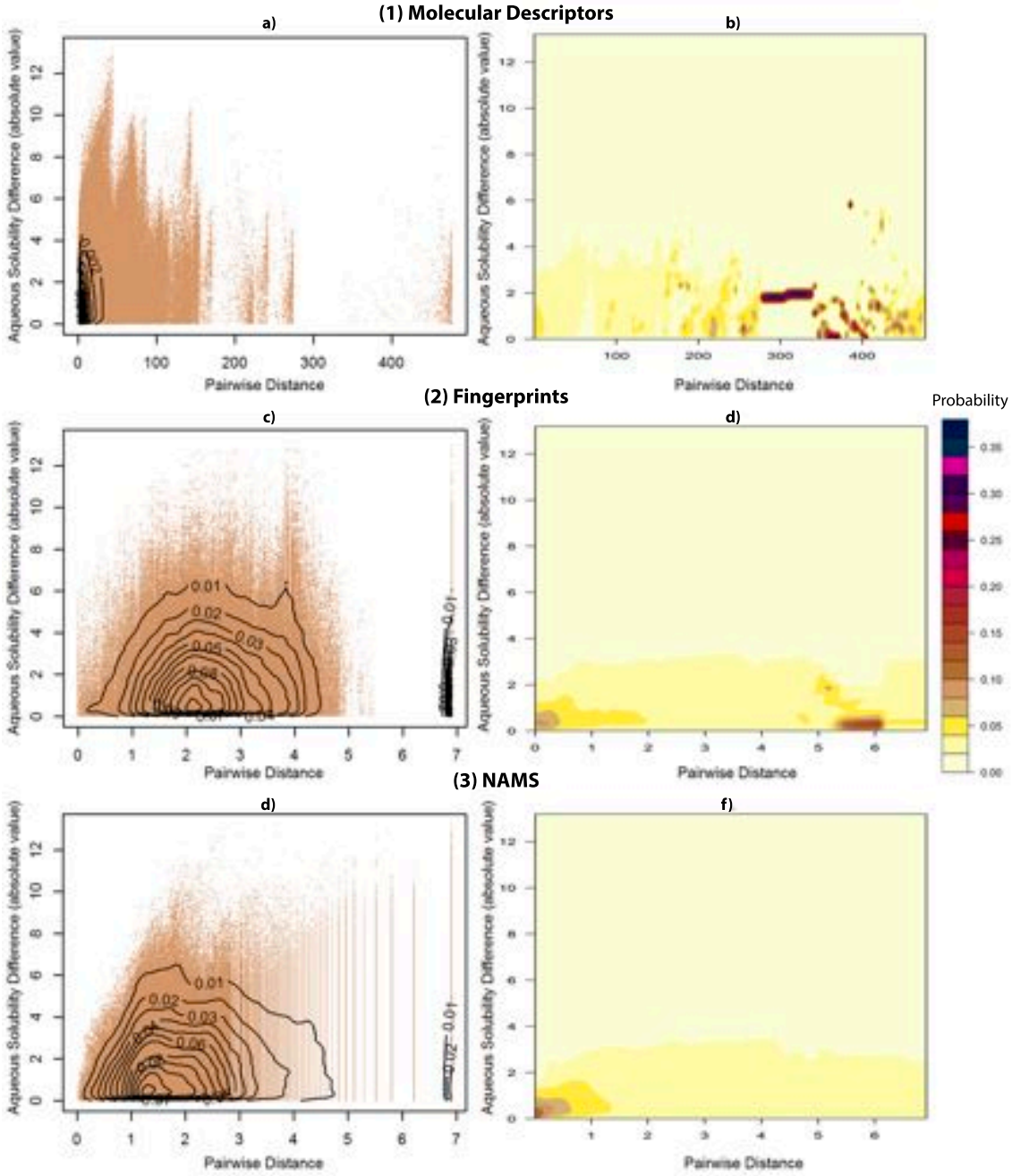


Figure 5.1: On the left side the plots represent the distribution of pairwise distance between pairs of compounds (training set B), calculated using *a)* molecular descriptors, *c)* fingerprints and *e)* NAMS and the corresponding absolute difference in the aqueous solubility value. The contour lines represent two-dimensional kernel density of the pairwise distance between pairs of compounds and the respective absolute difference in the aqueous solubility value. On the right side the plots show at each level of pairwise distance, using the *b)* molecular descriptors, *d)* fingerprints and *f)* NAMS for such calculation, the probability of finding a pair of compounds with a certain absolute difference in the aqueous solubility value.

Table 5.3: Summary of the best results (leave-one-out cross validation) obtained for training and testing sets (dataset B) using each dissimilarity matrix.

Molecular Representation	Method	*neighs*†	Type	nł	RMSE	%[-1.0, 1.0] ‡	q^2_{LOO}/Q^2
Molecular Descriptors	*DistKrig*	300	Train*	1033	0.9475	76.09	0.7840
			Test 1	258	0.8678	78.21	0.8143
			Test 2	21	0.9105	72.33	0.7496
Fingerprints	*CoordKrig*	8	Train*	1033	1.2407	65.27	0.6296
			Test 1	258	1.1161	68.34	0.6929
			Test 2	21	**0.7871**	**77.48**	**0.8129**
NAMS	*DistKrig*	5	Train*	1033	**0.7793**	**82.44**	**0.8537**
			Test 1	258	**0.8332**	**81.55**	**0.8288**
			Test 2	21	1.0941	73.18	0.6384

†Number of selected neighbouring molecules for each prediction
łTotal number of compounds in the set
‡% of predictions with error between -1.0 and 1.0
* Leave-one-out cross-validated results

Table 5.3 summarizes the best results for the training (obtained with leave-one-out cross-validation) and testing sets with each dissimilarity matrix selected based on the RMSE. The method used to select neighbourhoods was the moving neighbours with a fixed number of compounds. Detailed results are provided in Appendix C.5.

The best model for the training set was obtained using *DistKrig* coupled with NAMS to calculate the distance between molecules and reached a RMSE of 0.7796 which corresponds to a q^2 of 0.8537 with 82.44% of the compounds being predicted with an absolute error smaller than ±1 (Table 5.3). These results were obtained excluding one compound at the time for testing and using the most similar 5 compounds in the training set to predict its property. The results using NAMS to calculate the distance between the compounds tend to decrease with the increase of the number of compounds used to predict the property which complies with the similarity principle and also due to the high redundancy between the selected compounds interfering in the spatial correlation needed to construct the semivariogram. For testing set 1 and using the most similar 5 compounds in the training set to predict the property value, a RMSE of 0.8332 was obtained which corresponds to a Q^2 of 0.8288 with 81.55% of the compounds being predicted with an absolute error smaller than ±1 (Table 5.3). When evaluating the model

with testing set 2 and using the most similar 5 compounds in the training set to predict the property value, a RMSE of 1.0941 was obtained which corresponds to a Q^2 of 0.6384 with 73.18% of the compounds being predicted with an absolute error smaller than ±1 (Table 5.3). For test set 2, there are two compounds with large errors which heavily penalize the predictive scores of this model due to the small number of compounds in the set.

The results using *DistKrig* coupled with molecular descriptors to calculate the distance between the compounds tend to improve with the number of compounds used to predict the property, which complies with the preliminary analysis of the relationship between structural similarity and property difference, since there is an approximately inverse relationship between the pairwise distance and the aqueous solubility difference, especially for smaller distances since there is no guarantee that molecular descriptors that contribute most to the calculation of the distance between molecules represent the substructures that most influence the property value, thus justifying the fact that using only the most similar compounds would not be enough, leading to semivariograms that are too irregular to be fitted with a linear function by not showing any spatial correlation.

The prediction results using *CoordKrig* coupled with fingerprints to calculate the distance between the compounds tend to improve until 8 compounds are used, from this point on this tendency reverts due to the high redundancy and distance between the selected compounds. Due to the high redundancy between the compounds in the dataset when compared using fingerprints and to the existence of several pairs of structurally different compounds in the fingerprint-distance matrix with a score of zero, contrarily to the other distance methods *DistKrig* obtains low predictive power and in some cases it is even impossible to apply it due to ill-conditioned distance matrices with a large condition number which means that such matrix is almost singular and the computation of its inverse is not possible or it is prone to large numerical errors. The best predictive performance for the test set 2 was obtained with fingerprints (Table 5.3), since there are some compounds in the testing set that are significantly different from all compounds in the training set and which benefit with a less granular similarity score.

Table 5.4: Comparison of the predictive power of the model developed in this study with other published models with the best results (selected by the performance on the training set) for the same dataset (with different partitions of the data into training and testing) by Multi-linear Regression (MLR) Analysis and Artificial Neural Network (ANN) models and using different selections of molecular descriptors.

Reference	Model	Train set			Test set 1			Test set 2		
		nł	q^2/R^2	RMSE	nł	Q^2	RMSE	nł	Q^2	RMSE
Our model	Kriging	1033	0.85*	0.78*	258	0.83	0.83	21	0.64	1.09
Hou *et al.* (2003)	MLR	878	0.92	0.59	412	0.90	0.63	21	0.88	0.84
Yan & Gasteiger (2003)	MLR	797	0.79	0.93	496	0.82	0.79	21	0.56	1.20
	ANN	797	0.93	0.5	496	0.92	0.59	21	0.85	0.77
Liu & So (2001)	ANN	1033	0.86*	0.70*	258	0.86	0.70	21	0.79	0.91
Tetko *et al.* (2001)	MLR	879	0.86	0.75	412	0.85	0.81	21	0.77	0.99
	ANN	879	0.93	0.53	412	0.90	0.66	21	0.89	0.67
Huuskonen (2000)	MLR	884	0.89	0.67	413	0.88	0.71	21	0.83	0.88
	ANN	884	0.94	0.47	413	0.92	0.60	21	0.91	0.63

łTotal number of compounds in the set
* Leave-one-out cross-validated results.

Table 5.4 shows a summary of the best models found in the literature using dataset B, which cannot be directly compared due to the fact that different partitions of the data and different validation methods are used in each study. However, it is possible to observe that the predictive performance of *DistKrig* coupled with NAMS is within the range of the performances obtained by the best models in the literature for this dataset, especially when compared with Liu & So (2001) which uses the same partitions in training and testing sets and the same validation method.

5.4.3 Case C - Predicting Dihydrofolate reductase (DHFR) inhibitors activity

As a preliminary analysis, it was evaluated, in light of the structural similarity principle, the capacity of each of the structural similarity methods in study to discriminate molecules with different activity value solely based on their structural distance.

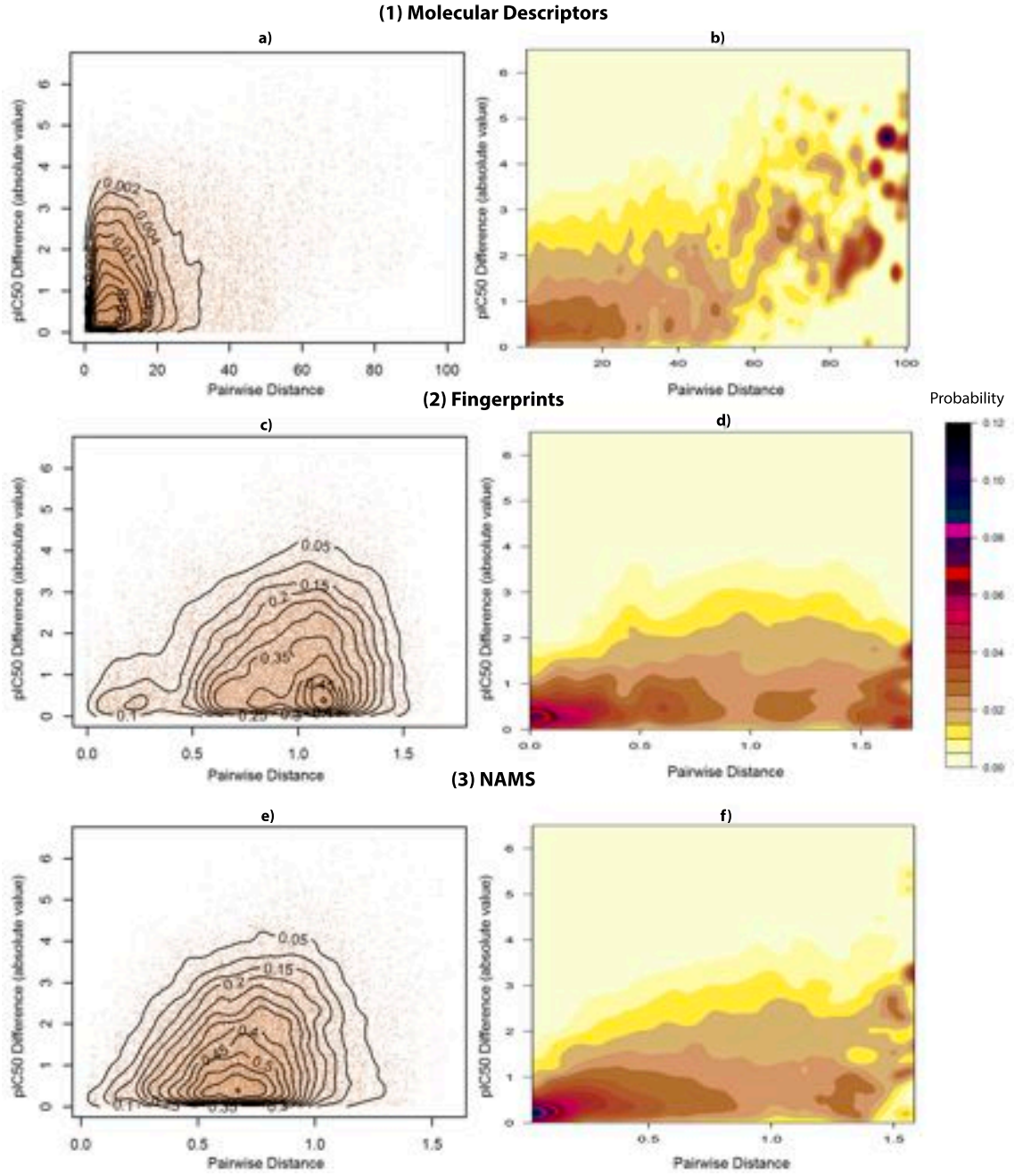


Figure 5.2: On the left side the plots represent the distribution of pairwise distance between pairs of compounds (training set A), calculated using *a)* molecular descriptors, *c)* fingerprints and *e)* NAMS and the corresponding absolute difference in the pIC_{50} value. The contour lines represent two-dimensional kernel density of the pairwise distance between pairs of compounds and the respective absolute difference in the pIC_{50} value. On the right side the plots show at each level of pairwise distance, using the *b)* molecular descriptors, *d)* fingerprints and *f)* NAMS for such calculation, the probability of finding a pair of compounds with a certain absolute difference in the pIC_{50} value.

Figure 5.2 displays plots showing on the left side the distribution of the pairwise distance between the 237 compounds in the training set, totalizing 27966 unique pairs of structures (excluding self-distances), calculated using a) molecular descriptors, c) fingerprints and e) NAMS and the corresponding absolute difference in the pIC_{50} value. While on the right side, Figure 5.2 displays plots showing at each level of pairwise distance, using the b) molecular descriptors, d) fingerprints and f) Noncontiguous Atom Matching Structural Similarity (NAMS) for such calculation, the probability of finding a pair of compounds with a certain absolute difference in the pIC_{50} value.

In general, it is possible to verify that both fingerprints and especially NAMS are able to discriminate the compounds in the training set according to their pairwise distance versus their difference in the pIC_{50} value, especially for most similar pairs of compounds, verifying the similarity principle (Johnson & Maggiora, 1990), since the plot of pairwise distance values versus absolute difference in the pIC_{50} values (Figure 5.2 - c) and e)) for the set of molecules exhibit a trapezoidal distribution, revealing a neighborhood behavior with a low frequency of pairs in the upper left triangle (very similar compounds with a high difference in the property value). In the probability distribution plot, it can be observed (Figure 5.2 - d) and f)) that for both Fingerprints and especially for NAMS there is an high probability for compounds that are very close to each other to have a small difference in the property value. Nevertheless, fingerprints have a higher number of similar pairs of compounds (even 100% similar) with a higher difference in the property values than NAMS which is contrary to the similarity principle. When using molecular descriptors, the relationship between the pairwise distance of the compounds and their difference in the property is less clear, even though there is a tendency for pairs of very dissimilar compounds have higher differences in the property value (Figure 5.2 - a) and b)). Therefore fingerprints and NAMS are more likely to obtain a discriminating metric space to be interpolated by kriging.

One problem in retrospective QSAR studies is that data is randomly sampled so that virtually all scaffolds are represented in training and test sets, being limited when new compound scaffolds appear or when the structure of a test compound varies significantly. To overcome this situation, we also performed a validation of the method in a real-world context of drug discovery using a

Table 5.5: Summary of the best results (leave-one-out cross validation) obtained for training and testing sets (dataset C) using each dissimilarity matrix.

Molecular Representation	Method	*neighs*†	Type	nł	RMSE	%[-1.0, 1.0] ‡	q^2_{LOO}/Q^2
Molecular Descriptors	*DistKrig*	236	Train*	237	0.8430	77.87	0.5564
			Test	124	0.9584	67.52	0.5043
Fingerprints	*CoordKrig*	10	Train*	237	0.8273	77.84	0.5728
			Test	124	0.9696	69.20	0.4926
NAMS	*DistKrig*	5	Train*	237	**0.8105**	**79.76**	**0.5900**
			Test	124	**0.8609**	**73.41**	**0.6000**
NAMS (**Temporal**)	*DistKrig*	5	Train*,§	313	0.8738	76.39	0.6535
			Test¶	84	0.8940	72.35	0.6163

†Number of selected neighbouring molecules for each prediction
łTotal number of compounds in the set
‡% of predictions with error between -1.0 and 1.0
* Leave-one-out cross-validated results
§Property measurements published between 1991 and 1998
¶Property measurements published between 1999 and 2002

temporal selection of training data and test. That is, using earlier published property measurements (1991-1998) as training data to predict later property measurements (1999-2002).

Table 5.5 summarizes the best results for the training (obtained with leave-one-out cross-validation) and testing sets with each dissimilarity matrix, selected based on the RMSE as well as results for temporal data selection using the best model settings. The method used to select neighbourhoods was the moving neighbours with a fixed number of compounds. Detailed results are provided in Appendix C.5.

The best model for the training set was obtained using NAMS to calculate the distance between molecules coupled with the method *DistKrig* and reached a leave-one-out cross-validated RMSE of 0.8105 which corresponds to a q^2_{LOO} of 0.5900 with 79.76% of the compounds being predicted with an absolute error smaller than ±1 (Table 5.5). These results were obtained excluding one compound at the time for testing and using the most similar 5 compounds in the training set to predict its property. The results using NAMS to calculate the distance between the compounds tend to decrease with the increase of the number of compounds used to predict the property which complies with the similarity principle and also due to the high redundancy between the selected compounds

interfering in the spatial correlation needed to construct the semivariogram. For the testing set and using the most similar 5 compounds in the training set to predict the property value, a RMSE of 0.8609 was obtained which corresponds to a Q^2 of 0.6000 with 73.41% of the compounds being predicted with an absolute error smaller than ± 1 (Table 5.5).

The prediction results using molecular descriptors to calculate the distance between the compounds tend to improve with the number of compounds used to predict the property, which complies with the preliminary analysis of molecular similarity, since there is not a clear relationship between the pairwise distance and the pIC_{50} difference, especially for smaller distances considering that there is no guarantee that molecular descriptors that contribute most to the calculation of the distance between molecules represent the substructures that most influence the property value, thus justifying the fact that using only the most similar compounds would not be enough, leading to semivariograms that are too irregular to be fitted with a linear function by not showing any spatial correlation.

The prediction results using fingerprints to calculate the distance between the compounds has a tendency to improve until 10 compounds are used, yet, from this point on this tendency reverts due to the high redundancy between the selected compounds interfering in the spatial correlation needed to construct the semivariogram.

For this molecular representation, the implementation *CoordKrig* that preliminarily transforms the distances between molecules into 2D coordinates, jitters duplicated coordinates and uses a spherical function to fit the semivariogram showed advantages, which may be related to the existence of several pairs of structurally different compounds in the fingerprint-distance matrix with a score of zero (corresponding to 100 % structurally similar compounds).

As already mentioned in the data description there were 36 inactive compounds with IC_{50} that have not been experimentally determined ($> 10\ \mu M$ and artificially labelled with an observed value of 3.30) that were not included in the training or testing sets. The settings of the best model (NAMS coupled with *DistKrig* selecting the 5 most similar molecules) were used for predicting the pIC_{50} of these inactive compounds considering a threshold of 6.0 for discrimination between highly active and inactive compounds as previously used in other studies

(Sutherland *et al.*, 2003, 2004). The mean prediction property for the inactive set is 4.22 ± 0.99 and only one inactive compound (id: *1-127977*) is predicted with a higher value (7.7) than the defined threshold.

Table 5.5 also shows the results obtained applying temporal selection to divide the compounds in training and testing sets using the best model settings obtained with random data selection. This model used NAMS to calculate the distance between molecules coupled with the method *DistKrig* and reached for the training set (experimental measurements obtained between 1991 and 1998) a leave-one-out cross-validated RMSE of 0.8738 which corresponds to a q^2_{LOO} of 0.6535 with 76.39% of the compounds being predicted with an absolute error smaller than ± 1. For the testing set and using the most similar 5 compounds in the training set to predict the property value, a RMSE of 0.8940 was obtained which corresponds to a Q^2 of 0.6163 with 72.35% of the compounds being predicted with an absolute error smaller than ± 1. Although the results obtained using temporal selection cannot be directly compared with the previously obtained using random selection, it is possible to observe that these are of comparable quality in terms of predictive performance.

Although the main objective of this study is not to compare the predictive performance of the proposed methodology with the state-of-art QSPR/QSAR approaches, it is important to have a general idea of the best results obtained for the dataset. Table 5.6 shows a summary of the best models found in the literature using dataset C, which cannot be directly compared due to the fact that different partitions of the data and different validation methods are used in the different studies. However, it is possible to observe that the best results of other authors for the test set are of comparable quality to the use of kriging coupled with NAMS. This approach also obtains better predictive results than the hybrid approach using variable importance as calculated by random forests (RF-VI) presented earlier in this document.

5.5 Discussion

Based on the results for all datasets in study, *DistKrig* coupled with NAMS for calculating the distance between the molecules is the natural choice for a

Table 5.6: Comparison of the predictive power of the models developed in this study - instance-based learning using structural similarity and Kriging and model-based learning hybrid approach using variable importance as calculated by random forests (RF-VI) and support vector machines to select 95 descriptors from descriptor sets A and D with other published models with the best results (selected by the performance on the training set) based on same dataset (with different partitions of the data into training and testing) by Comparative molecular field analysis (CoMFA) with partial charge calculation method MMFF94, 3D pharmacophores QSAR with Self-Consistent atomic Property Fields by Optimization (SCAPFOld), Hologram QSAR coupled with Partial Least Squares (PLS) and 2.5D descriptors coupled with Neural Network (NN) models.

Reference	**Model**	**Train set**			**Test set**			**Inactive set**	
		nł	q^2	RMSE	nł	Q^2	RMSE	nł	% inactives
Our model	Kriging	237	0.59*	0.81*	124	0.60	0.86	36	97
Our model	RF-VI	237	0.68†	0.72†	124	0.53	0.96	36	92
Mittal *et al.* (2009)	CoMFA	397	0.69‡	-	-	-	-	-	-
Totrov (2008)	SCAPFOld	-	-	-	124	0.64	0.84	-	-
Sutherland *et al.* (2004)	HQSAR-PLS	237	0.69†	0.71†	124	0.63	0.84	36	92
	2.5D Descp-NN	237	0.61†	0.79†	124	0.42	1.05	36	83

łTotal number of compounds in the set
*Leave-one-out cross-validated results.
‡Leave-several-out cross-validated results.
†10-fold cross-validated results.

final model, since for these datasets it is able to produce models with the best predictive performance using the smallest number of compounds needed to predict the property of interest. The obtained models are robust since similar predictive performances were obtained for both training and test sets. Also, the predictive performance of the models comply with the results obtained in the literature using typical QSPR/QSAR approaches, however two of the great advantages of kriging are the exploration of a richer hypothesis space by creating local approximations for each test instance and the estimation of the prediction error for each predicted value in contrast to typical model-based approaches that commit to a single global hypothesis that covers the entire instance space and estimate only a global model prediction error. In this section the neighbourhood selection strategies, prediction error, its relationship with the similarity between the compounds, the kriging estimated variance and its relationship with the true prediction error and the effect of the size of the training set in the predictive performance of the method will be analysed in detail for the best model obtained for each dataset.

5.5.1 Neighbourhood selection strategies

The results obtained in the study of different neighbourhood selection strategies showed that the robustness of the Kriging model with respect to errors in the predicted property data can be negatively influenced when the neighbourhood is large since there is an high degree of redundancy. Furthermore, redundant neighbourhoods will lead to ill conditioned matrices that increase numerical inaccuracies and, in many cases, non-invertible semivariance matrices.

The strategy that yields better predictive results employs moving neighbourhoods with a fixed number N of compounds. Therefore, this strategy was applied for prediction tasks in the subsequent models. For each dataset the number N of compounds should be optimized, however it was found that the optimal number is, usually, between 5 and 20 compounds. The combination of other strategies, such as deletion of negative weights or definition of a minimum distance between compounds, with moving neighbourhoods did not show any predictive advantage
.

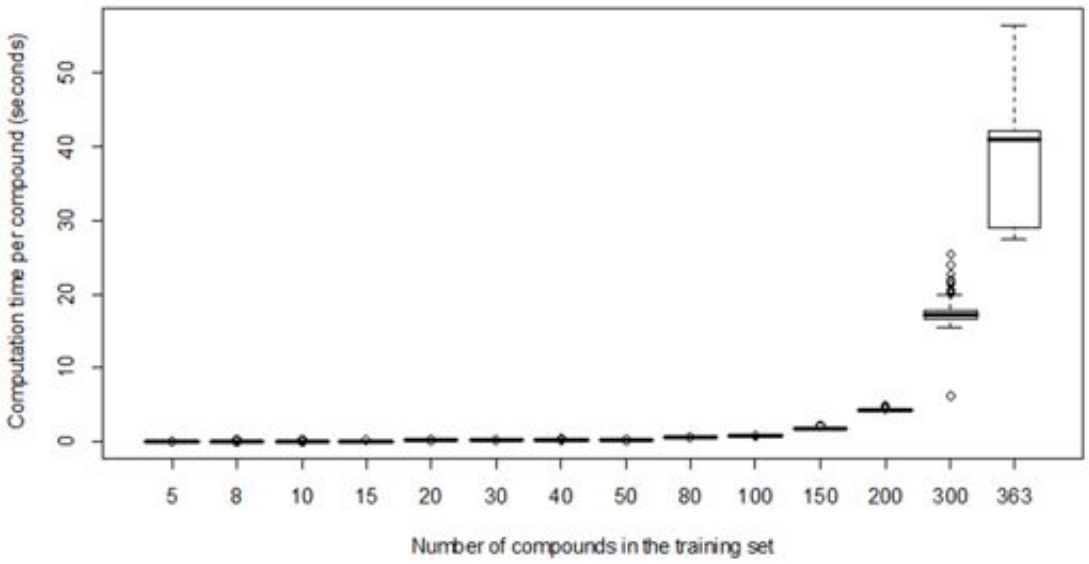
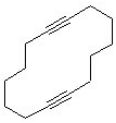
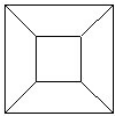
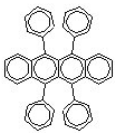
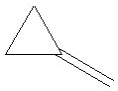
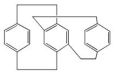
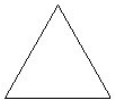


Figure 5.3: Relationship between the selected neighbourhood size and the calculation time for predicting enthalpy of formation of gas phase for each compound in the training set A1.

In addition to the predictive ability, a second reason for selecting a smaller neighbourhood is the reduction in time necessary to fit the Kriging model. This is an important issue as its time-consumption is generally regarded as one of the main drawbacks of the Kriging method (Rennen, 2009). The most computationally expensive step is the inversion of the correlation matrix as this requires large part of the total computation time and memory capacity and is thus worthwhile to reducing. Figure 5.3 shows the relationship between the neighbourhood size and the calculation time for each prediction. These calculations were performed in a standard desktop personal computer (CPU Intel Core i7, running at 2.0 GHz with 8 GB of RAM). It is clear that the computational time grows in an exponential way as the neighbourhood size increases. Besides improving the predictive ability of the model, smaller neighbourhoods will require less time either during training or in the test phase. This point is specially important in the situation where libraries of millions of compounds are being used or where the model is used for online monitoring and optimization.

5.5.2 Prediction error analysis

Table 5.7: Summary of the compounds with the highest prediction errors using the best model to predict enthalpy of formation of gas phase (case-study A1). It is also presented the minimum, average and maximum similarity between the test and training compounds.

ID	Compound Name	Min Sim (%)	Avg Sim (%)	Max Sim (%)	Obs Prop	Pred Prop	Pred Error	Molecular Structure
CO01875	1,8-Cyclotetradecadiyne	50.6	62.4	74.4	313.8	-117.0	430.8	
CO02606	Cubane	23.4	30.0	57.5	622.1	257.2	364.9	
CO02831	Tetraphenylnaphthacene	24.4	32.0	54.5	780.9	453.7	327.2	
CO01284	Methylenecyclopropane	38.7	48.7	70.0	200.5	513.0	-312.5	
CO02953	Pentacyclohexacosanonene	28.7	37.0	61.6	409.5	132.2	277.3	
CO01264	Cyclopropane	38.7	46.6	65.4	53.3	-167.3	220.6	

The predictive results obtained for case-study A1 are similar to the previously obtained using model-based approaches. The highest prediction errors were further analysed and are represented in Table 5.7. Again, it is possible to observe that prediction errors are higher for compounds with triple bonds or multiple cycles. However, this approach enables the analysis of the similarity between these compounds and the training set. The average similarity between test and training compounds is low and none of these six compounds have at least one compound in the training set that scores close to 85% (maximum similarity) which is the standard threshold to consider that two compounds are similar (Martin *et al.*, 2002). The similarity-based distribution of the neighbourhood of these compounds is a

good indicator of the high prediction error when applying this model, as they can be considered out of the domain of applicability of the training set. Excluding these six compounds of the set, yields much higher cross-validated performance measures namely, an RMSE of 38.2 and a q^2cv of 0.956.

The prediction errors obtained for the train and test sets using NAMS to calculate the similarity between the compounds for both datasets B and C were further analysed and are represented in Figure 5.4.

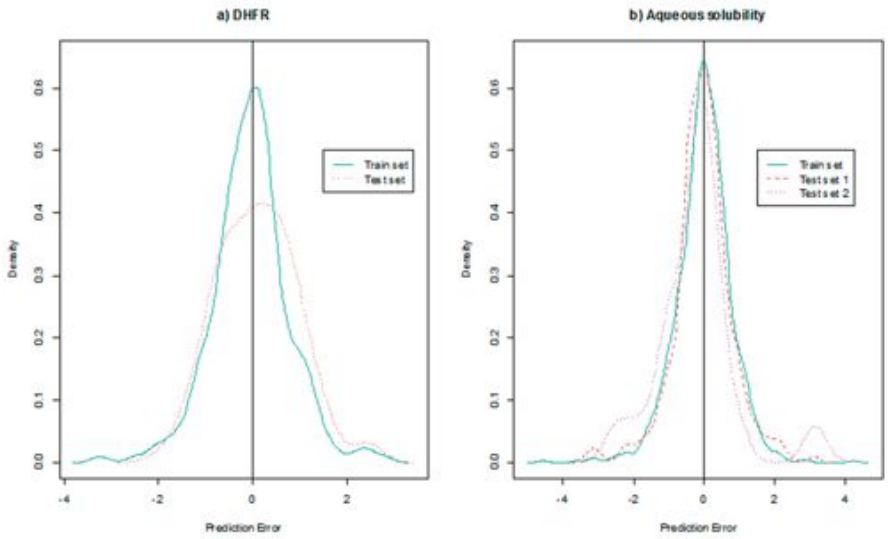


Figure 5.4: Density plots of the differences between the observed values and the predicted values for the train and test sets compounds using *DistKrig* coupled with NAMS to calculate the distance between compounds for the datasets a) DHFR and b) aqueous solubility.

For dataset B - aqueous solubility (Figure 5.4 - b)), *DistKrig* coupled with NAMS model is predicting the aqueous solubility for training set, test set 1 and 2 with the most probable errors -0.266, -0.008 and -0.070, respectively. For the training and test 1 sets, the prediction error has a narrowed density curve, condensing 82.4% and 81.6% of the errors between -1.0 and 1.0, explaining its performance in relation to the distance matrices. However, for test set 2 only 73.2% of the prediction errors are between -1.0 and 1.0 and there are two peaks in the density curve on both sides representing high positive or negative errors, demonstrating that their decrease in the predictive performance in relation to the test

set 1 is due to few compounds that are predicted with an high error.

For dataset C - DHFR, the model is predicting the pIC_{50} with a slight right bias (higher values than expected) and the most probable error for both training and testing sets is 0.196. Although for both, train and test sets, most errors (79.8% and 73.4%, respectively) are within the range of error values between -1.0 and 1.0, there are some higher errors that have a heavy weight in the RMSE calculation. The compound (ID: *1-233903*) with the highest difference (-3.241) between the observed and the predicted property value was inspected. Using the original reference (Broughton & Queener, 1991) from which the value of the property for this compound was compiled (Sutherland *et al.*, 2004), a mistake was discovered since the original value of the IC_{50} is 220 nM and not 220 μM, corresponding then to a pIC_{50} of 6.657 instead of 3.658. Therefore, the real difference between the observed and the predicted property value should be -0.242 instead of -3.241. Considering the heavy weight of this prediction error and all predictions that were affected with the wrong value of the property of this compound, we can advocate that the RMSE would be significantly lower (0.773 instead of 0.811 for the training set and 0.859 instead of 0.861 for the test set). We can also underpin the use of this method in curating datasets by the analysis of the prediction error in comparison with the most similar compounds.

The compound (ID: *1-122870*) with the second higher difference (2.763) between the observed and predicted property value was then selected and analysed. Figure 5.5 depicts the structure of the test compound *1-122870* and the selected 5 training compounds. The pIC_{50} values of all training compounds are lower than the observed pIC_{50} of the test compound and as expected, the predicted property is closer to the property of the most similar compounds. Figure 5.5 also shows the structure of the family of the test compound, which compared to the structures of the training compounds leads us to the conclusion that the high similarity scores between training and test compounds are due to a high similarity in the radicals R1 and R2 and in the pyrimidine ring, however all these compounds lack the core quinazoline substructure (two fused six-membered simple aromatic rings: a benzene ring and a pyrimidine ring) which is determinant for the high potency of the test compound *1-122870*.

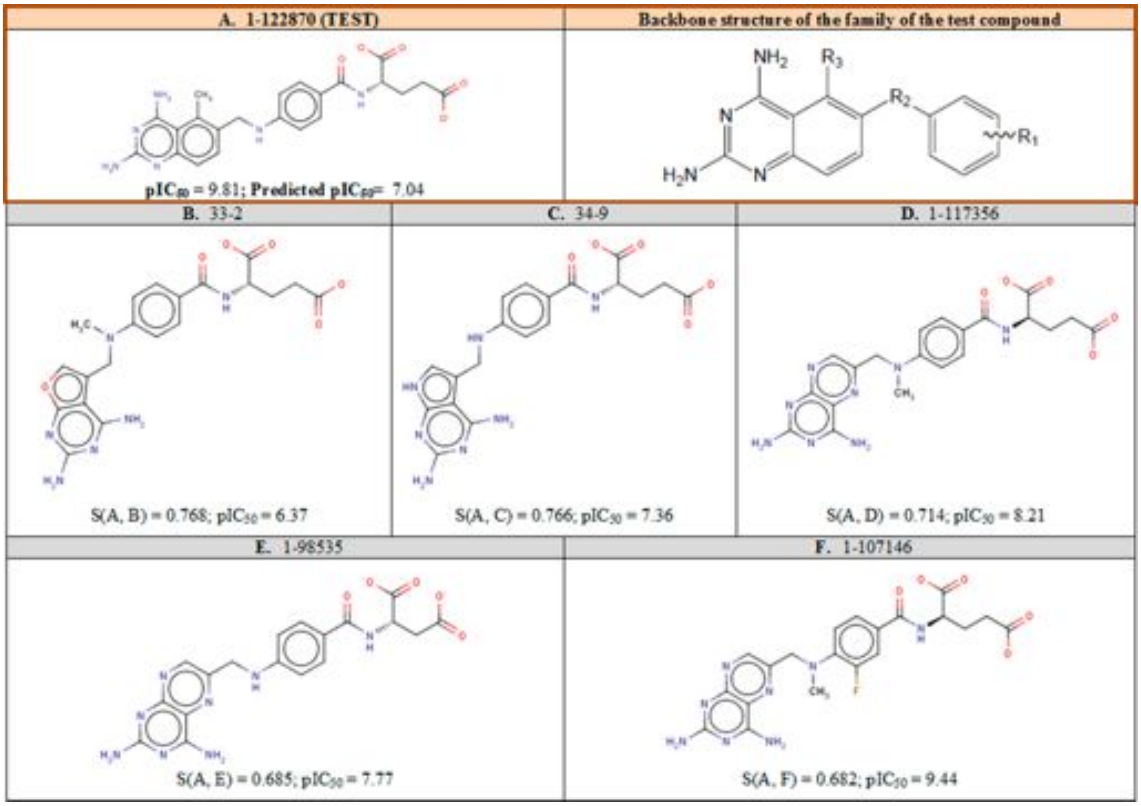


Figure 5.5: Structure of the test compound *1-122870* (***A***), the structure of the family of this compound and the 5 compounds selected for training (***B-F***). The similarity scores between the test and training compounds as well as the pIC_{50} values of each structure are also presented.

The results of the study of the temporal data selection for this dataset have confirmed the ability of the method demonstrated when applying random data selection. The training set was built using 26 different references while the test set was built using 11 different references, however the prediction errors do not show any trends based on reference or publication year. Most compounds with higher prediction errors are common using both random and temporal data selection.

5.5.3 Relationship between prediction error and molecular similarity

Figure 5.6 presents the relationship between the maximum similarity amongst the test compounds and the compounds of the training set that were selected for

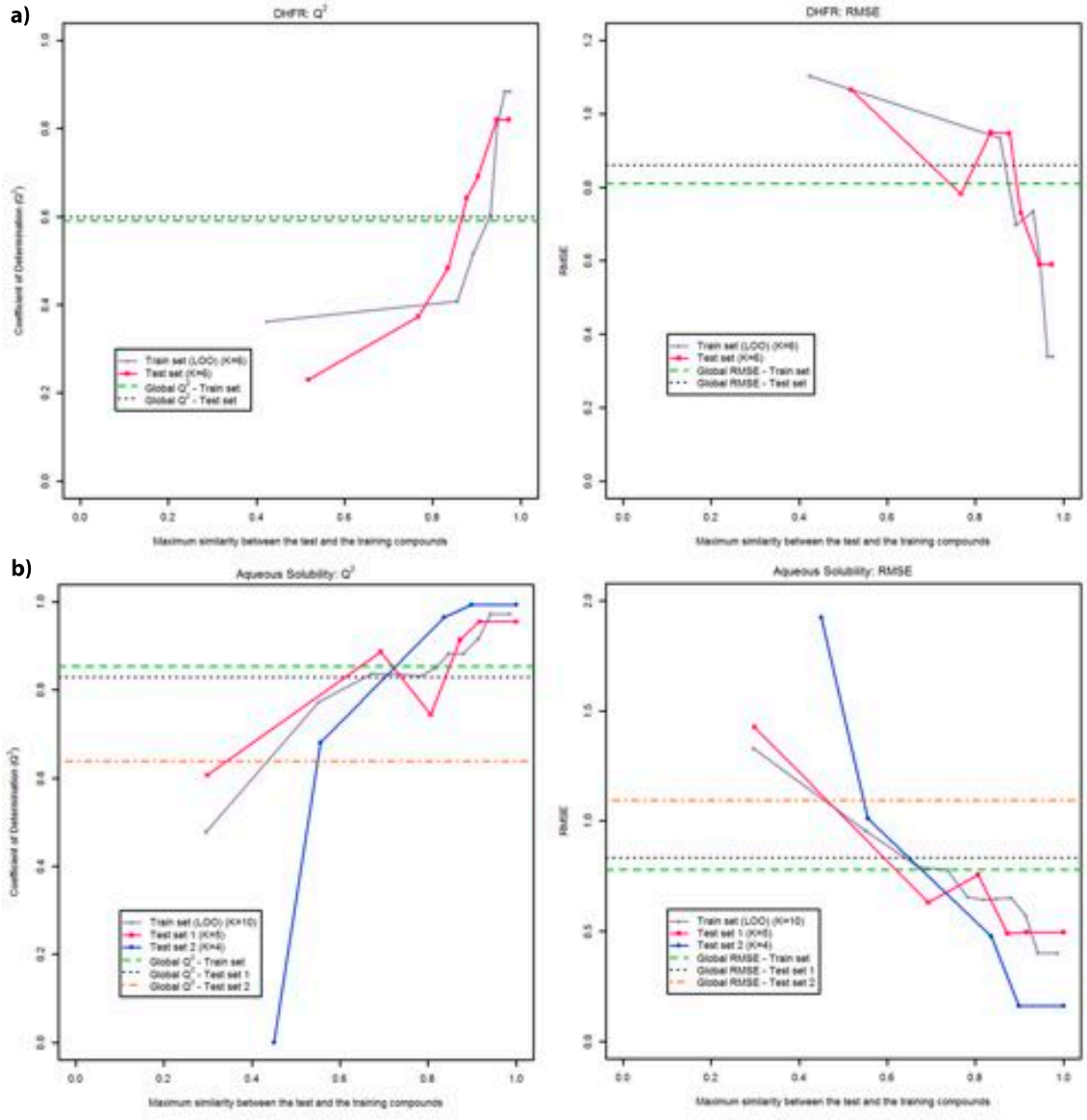


Figure 5.6: Plots showing the relationship between the maximum similarity between test compounds and compounds of the training set that were selected for predicting the property and predictive performance using Q^2 or RMSE for datasets B and C. The points marked in the lines represent the boundaries of the interval (K) for which the predictive performance metric is being averaged. The horizontal lines highlight the global RMSE or Q^2 obtained using all compounds. a) Dataset C: The similarity between the compounds and respective Q^2 or RMSE was averaged by 6 intervals containing 39 or 40 compounds each for training LOO and 20 or 21 compounds for testing. b) Dataset B: The similarity between the compounds and respective Q^2 or RMSE was averaged by 10 intervals containing 103 or 104 compounds each for training LOO, 5 intervals containing 51 or 52 compounds each for testing set 1 and 4 intervals containing 5 or 6 compounds each for testing set 2.

predicting the property averaged by intervals and respective predictive performance using Q^2 or RMSE for dataset B and C. In general it is possible to verify that Q^2 increases as the maximum similarity between the compounds increases and that RMSE decreases as the maximum similarity between the compounds increases.

Figure 5.6 - a) shows that if a threshold of 94% for the most similar compound to the test compound selected from the training set in dataset C is defined, which covers approximately 50% of all molecules, there is an high confidence in the predicted values with an expected $Q^2 > 80\%$ and RMSE < 0.59 which is significantly better than global Q^2 of 59% and 60% and RMSE of 0.81 and 0.86 for the training and testing sets respectively. It is important to highlight that for the last intervals of similarity the confidence in the results is very high, for example for the training set similarity scores range between 96.4% and 97.5% with an average Q^2 of 88.4% and RMSE of 0.34.

For dataset B (Figure 5.6 - b)) if a threshold of 84% for the most similar compound to the test compound selected from the training set is defined, which covers approximately 60% of the compounds, there is an high confidence in the predicted values with an expected $Q^2 > 85\%$ for the training set and test set 1 and 96% for the test set 2, and RMSE < 0.64 for the training set and 0.47 for the testing sets 1 and 2. Again it is important to emphasize the results for the last interval of similarity, for example for the test set 2 which *DistKrig* coupled with NAMS obtained worse results ($Q^2 = 64\%$ and RMSE $= 1.19$) than *CoordKrig* coupled with Fingerprints ($Q^2 = 81\%$ and RMSE $= 0.79$), however for the last interval of similarity scores which ranges between 89.9% and 99.0%, the average Q^2 is 99.0% and the RMSE is 0.16. These observations lead us to the conclusion that the existence of at least one compound in the training set that has a high similarity with the test compound allows making predictions with high confidence and minimal error. A complementary observation is that this method is able to identify regions of the molecular space that are lacking compounds with experimentally determined properties. These regions are ideal targets for experimental determination of properties of new molecules, which in turn will provide a broader coverage of the molecular space, resulting in a better model performance.

5.5.4 Kriging estimated variance and its relationship with prediction errors

The kriging estimated variance of each prediction depends on the arrangements of the observed values with respect to each other and with the location of the test compound in relation to the training compounds and it is completely independent of the true property of the test compound. Therefore, kriging provides the estimated variance at every estimated point, which is a great indicator of the accuracy of the estimated value and signs areas for which more experimental measures are needed.

In general there is a strong correlation between the kriging estimated error and the absolute true predicted error of each compound of dataset B and C. There are some cases for which the kriging estimated error is higher than the true prediction error which do not represent a problem, since it is in agreement with the notion of confidence interval, however there are also some cases that require further investigation since the kriging estimated error is smaller than the true prediction error. For that purpose, three compounds were identified from the test set C representing three different situations: **a)** a compound (ID: *1-351521*) which has a high true prediction error of 2.546 and a low kriging estimated error of 0.170; **b)** a compound (ID: *36-6e*) which has a low true prediction error of 0.649 and a high kriging estimated error of 1.834; and finally **c)** a compound (ID: *9-13e*) which has a low true prediction error of 0.007 and a low kriging estimated error of 0.093. Figure 5.7 shows the distance between the training and test compounds versus the pIC_{50} difference for each presented situation as well as the distribution of the properties in the compounds selected for training, the observed and predicted pIC_{50} of the test compound. Situation **a)** is represented in Figure 5.7 - a) which shows that the distance between the set of selected training compounds and the test compound *1-351521* is identical for all of them and relatively low (approximately 0.3). Also between the training compounds the property value is almost invariant. Therefore, this situation is translated in a low kriging estimated error, although due to small differences in key groups (similar to the case presented in Figure 5.5), the property of the test compound is significantly different from the properties observed for the training compounds.

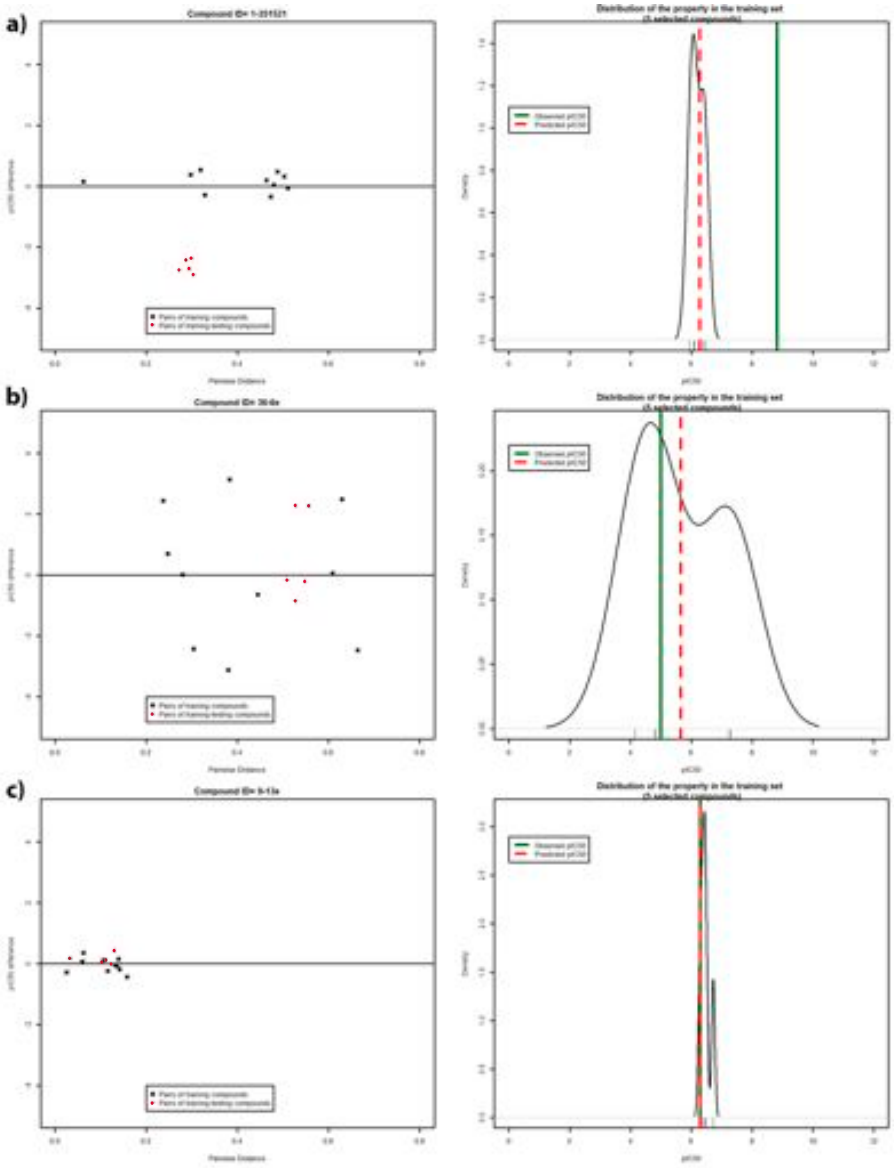


Figure 5.7: On the left side are represented the plots of the differences between the observed pIC_{50} values and the pairwise distance for all unique pairs between the 5 compounds selected for training (black points) and between the test compound (a)*1-351521*, b) *36-6e*, c)*9-13e*) and the training compounds (red points) in dataset C. On the right side are represented the density plots of the distribution of the property values in the selected training set (5 compounds) and the observed (green line) and predicted value (red dashed line) for the property of the test compound a)*1-351521*, b) *36-6e*, c)*9-13e* in dataset A.

The situation **b)** is represented in Figure 5.7 - b), which is not as fallacious as situation **a)**, since the true predicted error is within the estimated kriging error of [-1.834, 1.834]. The set of training compounds selected to predict the property of *36-6e* are relatively distant in relation to each other and to the test compound and the distribution of the property values in the compounds selected for train show an high range, thus an high estimated kriging error was expected. The situation **c)** is represented in Figure 5.7 - c) and it is an example of an ideal situation, since the true prediction error is within a narrow estimated kriging error interval of [-0.093, 0.093]. The set of training compounds used to predict the property of *9-13e* are similar to the test compound, have a low extent of distance scores as well as comparable differences in the property value.

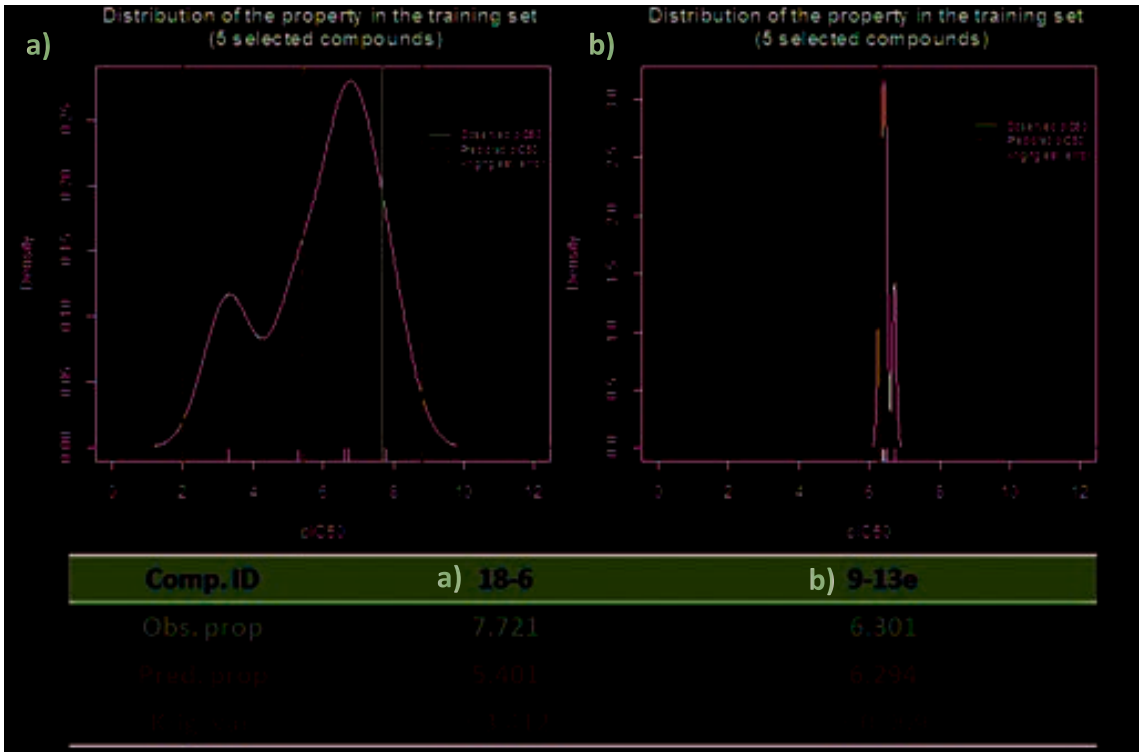


Figure 5.8: Relationship between true prediction error and estimated predicted error by kriging to assess the quality of prediction: a) high true prediction error and high estimated predicted error; and b) low true prediction error and low estimated predicted error.

When assessing the quality of a prediction, being able to accurately measure

its prediction error is of key importance. Models are not expected to accurately predict all properties for any possible compound, however the ability of a method to correctly predict this error provides more confidence and security in the resulting conclusions for each compound. Figure 5.8 presents two examples of predictions (case-study C): a) property of compound 18-6 is predicted with an high true predicted error, however within the margin of estimated predicted error by kriging which is wide, denoting lack of confidence in the predicted property; b) property of compound 9-13e is predicted with a low true predicted error and within a very narrowed margin of estimated predicted error by kriging denoting high confidence in the predicted property.

5.5.5 Effect of the training set size on the predictive results

The obtained models are limited in applicability by the data from which they are constructed, however this issue is seldomly addressed by reporting the kriging estimated variance of each prediction as a measure of extrapolation - high estimated kriging variances are obtained for compounds that are out of the applicability of the model by the lack of similar "neighbours", while low estimated kriging variances indicate that the model is able to predict that property value with high confidence. As already shown, the methodology provides improved predictive results with the increase of similarity between the training and test compounds. Based on these results, it is likely that large datasets will improve the predictive performance of the method, as the metric space has more instances the probability of finding training compounds that are more similar to the test compound increases. To test this hypothesis, the training set of dataset B (n = 1033) was used to create smaller datasets with n = {50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1033} compounds. Each of these subsets were created by random sampling n compounds 100 times. The predictive results (using RMSE) using each of these subsets to predict the aqueous solubility of the test set compounds (n = 258) are summarized in Figure 5.9 a). It is possible to observe that the predictive results increase as the number of compounds used as training set increases approximately following a power law distribution, however the gain in

the predictive results becomes asymptotically reduced with the increase of the number of compounds in the training set. As expected, in Figure 5.9 b) it is possible to observe that the maximum similarity between training neighbourhood and test compounds increases as the number of training compounds increases accompanying the improvement of the predictive results. It is important to note that this method always outputs the same predictive results as long as the same training set is used, as it can be observed when 1033 compounds are used for training.

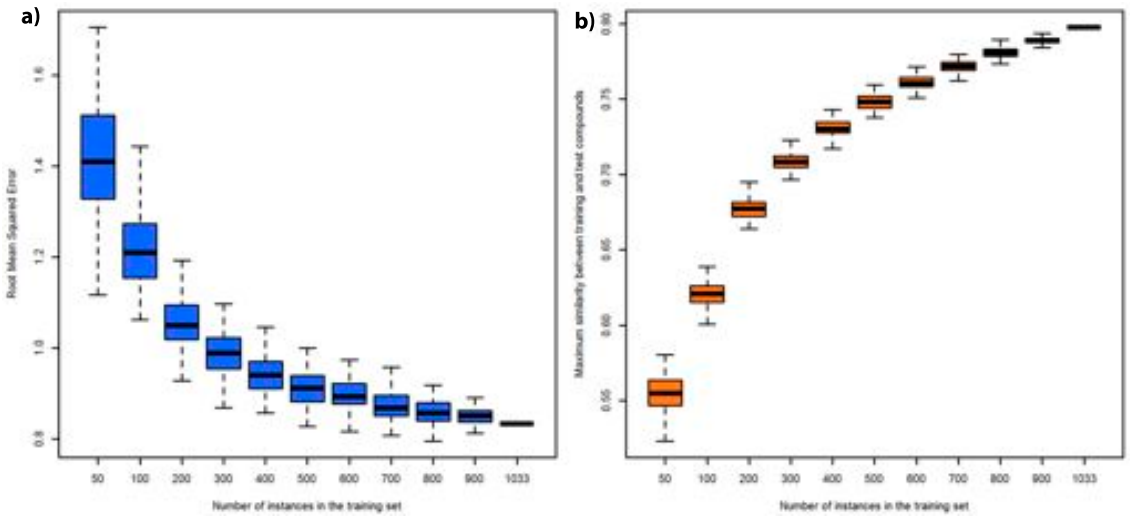


Figure 5.9: Effect of the training set (dataset B) size on **a)** test set predictive results using the Root Mean Squared Error (RMSE) and **b)** Maximum similarity between training neighbourhood and test compounds. Each of these training subsets were created by random sampling without replacement n compounds 100 times.

5.5.6 Assumptions and Limitations

Kriging is based on a statistical method which creates an interpolated map and output error map with the standard errors of the estimates, as such, the assumptions of the method should be considered carefully. The main assumption is stationarity (spatial homogeneity). If the change of data points from one neighbourhood to the next is too abrupt there may be discontinuities even though the actual phenomenon is continuous. If there is a spatial dependence between

points that are closer together, the semivariogram will have small semivariance and this semivariance is expected to increase with distance. If this assumption is held, just a few kriging model parameters have to be estimated from the data to make optimal predictions and valid statistical inferences. Therefore, the similarity metric used to map the compounds in the metric space should be able to discriminate the compounds according to the similarity principle. Furthermore, for the datasets in this study the assumption of being quasi-stationary does not apply to the entire dataset but only to the search neighbourhood under which the estimation model is fitted. Most of these sub-areas meet the local quasi-stationary assumption (homogeneity and density compromise) when analysing the pattern of the semivariogram cloud. No properties are guaranteed, when the wrong variogram is used, however, typically still a good interpolation is achieved even in cases of no spatial dependence in which the kriging interpolation is only as good as the arithmetic mean. The error map reflects data locations and it depends entirely on data configuration and semivariance function, therefore discontinuities will also be reflected in the kriging estimated variance. Furthermore, the use of small neighbourhoods is also advantageous in terms of computation since the basic math of this methodology involves the inverse of n by n matrix, where n is the number of data points used to predict the properties of a new compound.

As stated above this method, as most QSPR/QSAR methods, relies on the similarity property principle, which states that molecules that are structurally similar are likely to have similar properties. However, there are some exceptions to this similarity principle, most obviously in the case of activity cliffs where even a small structural change can be associated with a dramatic property shift. One of the advantages of OK in relation to other techniques based on distance (e.g. K-Nearest Neighbours) is that it considers not only the distance between the test and training compounds but also the distance between all training compounds. If the dataset is broad this problem is amended because the relationship between training compounds might indicate this discontinuity. The example depicted in Figure 5.10 shows a situation in which the most similar compounds have considerable differences in the property value, however the relationship between

all training compounds leads to a correct property interpolation of the test compound with a low absolute prediction error of 0.27, which is in agreement with the expected error of approximately 0.25 for the maximum similarity level between training and test compounds.

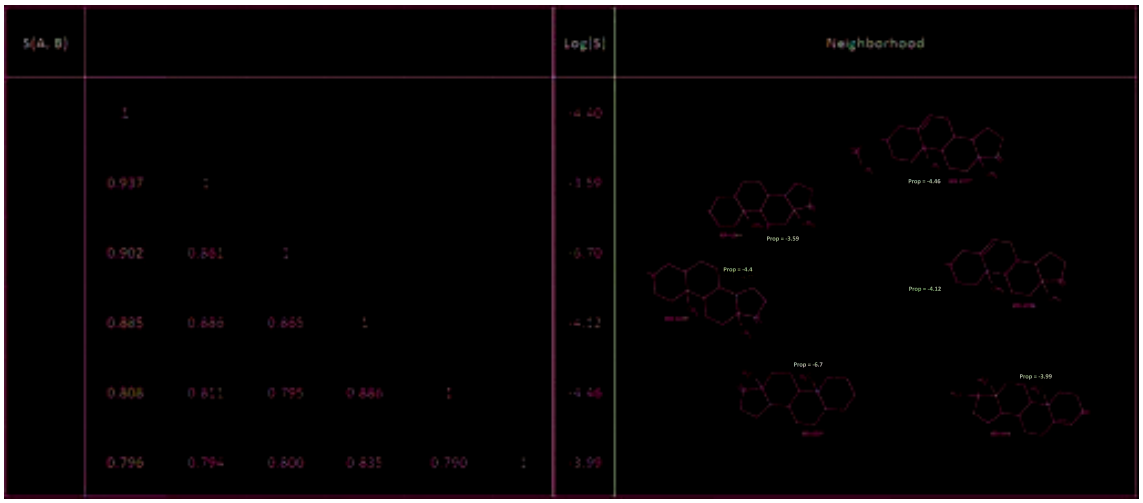


Figure 5.10: Example of a situation (data set B – aqueous solubility) in which the most similar compounds to the test compound (ID: 1259) have considerable differences in the property value. The relationship between training compounds leads to a correct property interpolation of the test compound with a low absolute prediction error of 0.27.

Nevertheless, there are some rare situations where the relationships between training compounds and test compound are not enough to correctly interpolate properties of compounds that have small structural differences associated with a dramatic property shift. An example of such situation is depicted in Figure 5.11, where the most similar compounds are all highly active and the test compound is highly similar, yet inactive. The property of this compound is predicted with an high prediction error of 4.40 when compared with the expected prediction error of approximately 0.81 for the maximum similarity level between training and test compounds. In such situations a specific weighting scheme for small structural modifications that are able to produce high property differences could be advantageous.

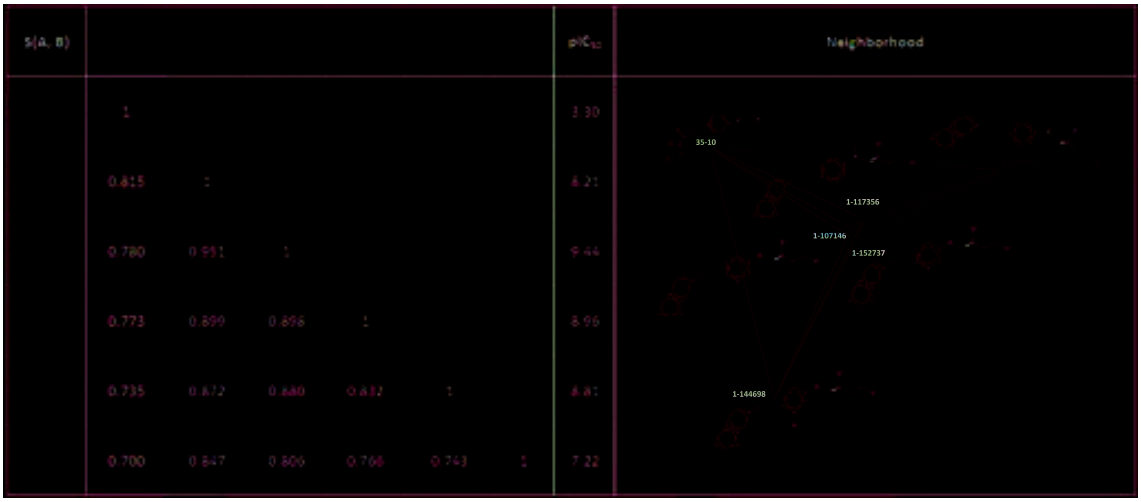


Figure 5.11: Example of a situation (data set C – DHFR inhibitors activity) in which the most similar compounds to the test compound (ID: 1-127977) are all highly active, yet the test compound is inactive. The relationship between training compounds does not leads to a correct property interpolation of the test compound (7.70) with an high prediction error of 4.40.

Even so, the principle is a very useful one for which there is substantial supporting evidence in the predictive results that were obtained and in large part these capabilities of the model are enhanced by the degree of structural similarity between the training-set and test-set molecules. Therefore, the quality and coverage of the training set is a key element for the predictive capabilities of the method.

5.6 Summary

In this study, we have proposed a new method for predicting chemical, physical or biological properties of chemical compounds solely based on structural similarity functions to define the chemical space that is explored by kriging models, in order to predict unmeasured properties of compounds.

The overall predictive results of the proposed methodology applied to three different datasets were good and within the confidence margins of other studies found in the literature or using model-based approaches, presented earlier in this document, using the same datasets. However, this new approach showed several

advantages relatively to other QSPR/QSAR approaches, namely (1) it makes use of the dissimilarity/similarity between the compounds and it is not necessary to use any feature selection and no prior knowledge of the problem or property is necessary, thus it may be applicable to most QSPR/QSAR studies directly and to any possible compound (even to compounds that were never synthesized) as long as its structure is known, (2) the similarity map that positions each molecule of the dataset in the chemical space can be used to predict any chemical or biological property of the compounds as long as experimental data is available, (3) it is possible to identify the situations for which prediction errors are deemed to be higher (measure of extrapolation) by estimating the kriging variance for each prediction - high estimated kriging variances are obtained for compounds that are out of the applicability of the model by the lack of similar neighbour compounds, while low estimated kriging variances indicate that the model is able to predict that property value with high confidence, (4) the model is readily understandable rather than a black box model, as it is possible to verify which of the compounds more directly impact the current estimation, (5) new compounds can be easily included as well as removed from the pool of training compounds since in this approach the target function is approximated locally for each test compound instead of an overall model that requires re-training each time the training set is changed, (6) the method can be applied to datasets of any size, however the predictive results are more likely to improve with the increase of the number of training instances as the probability of finding neighbour compounds with higher similarity increases, and (7) searches for the relationship among measured properties (richer hypothesis space) rather than approximate the modeled system by fitting the parameters of the selected basis functions (single hypothesis space). This approach can simultaneously solve multiple problems and deal successfully with changes in the problem domain.

The results showed that optimizing a fixed number of compounds that are closer to the test compound to determine its neighbourhood will minimize the prediction error in relation to the use of the whole training set. Besides improving the predictive results, the selection of neighbourhoods also reduces training and testing time, avoiding numerical inaccuracies and improving the robustness

and interpretability of the model. The results of the application of this methodology also showed that the existence of at least one compound in the training set that has a high similarity with the test compound allows making predictions with higher confidence and reduced error. Another important conclusion is that the current approach can be used to guide the extension of the training set and exploration of new promising regions within the molecular space by suggesting new molecules that can be used as seed compounds for experimental property determination, which in turn will improve the model quality by providing a broader coverage of the molecular space, as well as being used for dataset curation proposes by analysing the prediction error and the structure/property of the selected neighbour compounds.

The estimated variance resulting from kriging for each prediction showed a strong correlation with the true prediction error, which proves that it can be used as a quality measure of each kriging prediction since it provides a confidence interval in the predicted value. It is our conviction that kriging estimated variance can also be used to interactively determine the number of compounds that should be used to make a prediction based on the minimization of the kriging estimated variance.

For all datasets that were tested the similarity function NAMS to map the compounds in the chemical space using random or temporal data selection yields better validated predictions with a smaller number of compounds (as nearest neighbours) for each prediction than using molecular descriptors or fingerprints. NAMS yielding better results was expected since a preliminary analysis comparing the pairwise distance between the compounds and their property difference showed that NAMS was able to discriminate the compounds better in accordance with the similarity principle: similar compounds tend to have similar property values and *vice-versa*. The application of a feature selection technique prior to the calculation of molecular similarity using molecular descriptors could be advantageous, however the objective of the study was to build a universal map of structural relationships in the chemical space that was not dependent on the property in study. The analysis of the prediction errors showed that a similarity metric that would weight differently substructures that have a higher (positive

or negative) impact in the property value could improve the predictive performance of the method. However, any other similarity/dissimilarity approach can be applied in this methodology. In general, the predictive results are affected by redundancy between the compounds and by predictive maps that present patterns with several pairs of compounds at the same distance but with considerably different values of property and *vice-versa*.

This study was limited to predicting properties based solely on the structure of the compound as it may be used for any possible compound (even compounds that were never synthesized) and do not require any knowledge on their bioactivity or chemical/physical properties. Future work includes the development of a weighting schema in the similarity function to include both structural similarity and property profiles (using methods such as High Throughput Screening Fingerprints (Petrone *et al.*, 2012) or Similarity Ensemble Approaches (Keiser *et al.*, 2007) in order to accurately predict similar properties of compounds even when molecules are not structurally similar.

Chapter 6

Conclusions

"Every attempt to employ mathematical methods in the study of chemical questions must be considered profoundly irrational and contrary to the spirit of chemistry. If mathematical analysis should ever hold a prominent place in chemistry - an aberration which is happily almost impossible - it would occasion a rapid and widespread degeneration of that science."
$\sim$ Augustus Comte, *Philosophie Positive* (1830)

If Comte had lived long enough to see the development of chemistry (and more specifically chemoinformatics), he might have retracted these words. A dissenting and correct view:

"The more progress the chemical sciences make, the more they tend to enter the domain of mathematics, which is a kind of center to which they all converge. We may even judge the degree of perfection to which a science has arrived by the facility to which it may be submitted to calculation."
$\sim$ Adolphe Quetelet (1828)

6.1 Overall Approach

"All science is computer science". When a New York Times article in 2001 used this title, the general public was aware that the introduction of computers has changed the way the experimental sciences develop. There has never been a more pressing time in drug discovery and industrial production to reduce development time and cost. *In silico* techniques offer inexpensive methods to assist these processes. This thesis has covered various aspects of QSPR/QSAR from model

choice and optimisation, to descriptors and their selection, appropriate machine learning techniques to establish a relationship between structure-property and model validation and evaluation. Every choice made before building the model that establishes a relationship between structure-property is important as it affects the potential outcome and therefore usefulness of it. Machine learning offers a plethora of techniques to build a model upon, each with advantages and disadvantages. It is also important to guarantee that a reasonably sized data set is available and if the data set is to be split into train, hold and test sets, this requires careful handling to ensure the sets are equally distributed. Descriptors will need to be generated in order to describe the structure of molecules, however, there are thousands to pick from again, with advantages and disadvantages. Regardless, the design of QSPR/QSAR models for predicting many important physico, chemical and biological properties is still an important topic of research, because the process of QSPR/QSAR model design and evaluation is relatively complicated. Moreover, currently available methodologies have only limited functionality or are strictly aimed at some particular chemical problem. Furthermore, most of these methodologies work like "black boxes" without a detailed understanding of each prediction and expected prediction error. These facts are nowadays a strong bottleneck and new solutions for QSPR/QSAR modelling process are still demanded by the scientific community. We accepted this challenge and focused this research work on QSPR/QSAR modelling using machine learning methods. These motivations served as the basis for the following thesis previously defined that guided the research work detailed in this dissertation: *it is possible to improve the current models for the prediction of physical, chemical and biological properties based solely on the chemical structure using advanced automated analysis solutions based on Machine Learning.* The aims of the study covered the development and implementation of cutting-edge machine learning and statistical modelling algorithms for handling large-scale chemical data in order to improve the prediction of properties not only in terms of predictive power but also improving the robustness and comprehensibility of such methodologies. At the beginning, the theoretical basis of QSPR/QSAR modelling and the specificities of the representation of chemical structures in computer readable formats which are required for data analysis were studied. Physical, chemical as well as biological properties are in large part

determined by the molecular structure. There are several ways to represent a molecular structure and different representations contain different chemical information. One of the major tasks in automated extraction of meaning, patterns, and regularities using machine learning methods is to represent chemical structures, to transfer the various types of chemical information taking into account their complex and heterogeneous nature into a machine-readable representation that can be processed by a machine learning model. Hence, machine-readable representations and machine learning models that can handle and extract the right chemical data according to the chemical property that needs to be predicted were studied and selected. In parallel, we have studied currently available software tools and existing methodologies for QSPR/QSAR modelling. Based on this know-how and on the collection of several case-studies and respective datasets, a new methodology to select descriptors in QSPR/QSAR modelling was designed, developed, implemented, validated and evaluated. The number of different descriptors reaches thousands therefore one may be tempted to leave the descriptor selection process to algorithmic techniques. While this may lead to high accuracy of the model, the objective of this work is a comprehensive and interpretable model to give clear insight into the structure-property relationship. Furthermore, smaller models tend to generalize better than larger models, and tend to be statistically more robust. Therefore, after numerical descriptors have been calculated for each compound, its number should be reduced to a set of them that are information rich while being as small as possible. The developed approach uses random forests, not as modelling tools for themselves, but as a method capable of identifying the most important features of a given modelling problem, which are then used as input variables to Support Vector Machines models. It is important to note that random forests were the selected algorithm due to the enumerated advantages; however, in principle, any machine learning able to produce a ranking of variable importance could be applied. The second part of this hybrid algorithm uses a ranked list of variables, ranging from the most to the least important, to train SVM models using a stepwise approach of adding one variable for each model according to its predefined rank. Once again, it is important to note that, in principle, any non-linear machine learning method could be applied. The parameters of both models were optimized and the

effect of correlated variables studied. This methodology was successfully applied to both regression and classification problems and to predict different properties of datasets with different characteristics, namely standard molar enthalpy of formation of gas phase of a set of similar hydrocarbons, standard molar enthalpy of formation (gas, liquid and crystalline phases) and standard molar enthalpy of phase change (vaporization, sublimation and fusion) of an extensive (ranging from simple to complex structures) and highly diverse set of compounds, citotoxicity parameters of exposure to drugs at a population-level using a highly diverse set of pharmacological highly complex compounds and finally to predict if a compound is able to cross the Blood-Brain Barrier using an extensive and highly diverse set of compounds in a *real-world* drug discovery scenario. From the analysis of the obtained results for these QSPR/QSAR case-studies, we can conclude that the presented methodology performs well for high-dimensional data and it is robust even in the presence of highly correlated variables. The feature selection step yields lower prediction errors for all case-studies with a smaller number of variables in relation to the best models trained with all descriptors or using popular feature selection or dimensionality reduction techniques such as Genetic Algorithms or Principal Components Analysis. These reduced errors are relevant with significant chemical and economical importance, but they are also important in terms of computational performance since a smaller number of descriptors need to be calculated producing simpler models that are more robust and comprehensive. It is then safe to conclude that SVMs alone are not able to perform a good optimization, and by combining with a variable selection step we can obtain a minimum subset of important variables to train a faster and more robust model, yielding better prediction performance. Another objective of this work is to implement and assess some existing prediction models verifying the quality of the produced results as compared with the proposed methodologies. Some models described in the literature were presented and others were implemented such as the ELBA method. The implementation of the ELBA method gave an extensive insight on how to extract characteristics of a molecular structure, although its main disadvantage is that it is limited to hydrocarbons. The main conclusion of this first part of the research work is that the proposed methodology can be used to predict most physical, chemical and biological properties and for most types of

compounds as long as their structure is known, improving the prediction performance and robustness in relation to other QSAR/QSPR methodologies, providing faster and more cost-effective calculation of descriptors by reducing their number, and providing a better understanding of the underlying relationship between the molecular structure represented by descriptors and the property of interest. For one of the case-studies it was developed a database for structural and thermochemical properties of organic compounds and a Web interface to access the data and estimating properties. This information system (ThermInfo) is publicly available at http://www.therminfo.lasige.di.fc.ul.pt and it brings together critically evaluated values of thermochemical properties of pure substances and structural data. Undoubtedly experimental values are the "gold standard" to assess the quality of a prediction method, on the other hand, the accuracy of these methods is directly dependent on the size, completeness and accuracy of the testing/training data. Thus databases like the one provided by ThermInfo is absolutely critical to the development and application of properties prediction methods. ThermInfo also makes available features to predict thermochemical properties based on a user-friendly interface and accepting inputs in several formats. In general user's feedback about ThermInfo's usability is very positive. Overall, it is expected that ThermInfo will be a useful tool for the thermochemical research and industry community as well as educators, students and the general public.

In the second part of this work another avenue to explore the relationship between structure-property was proposed. Given that similar molecules tend to have similar properties, we alternatively developed a completely new approach in QSPR/QSAR modelling focused on instance-based machine learning for predicting properties of compounds using the similarity-based molecular space. However, the concept of molecular similarity is abstract, problem-dependent and subjective and its definition is, to a great extent, a semantic question. Molecular similarity depends on comparative perception without a defined standard. Various methods to define structural similarity between molecules available in the literature were studied, but so far none have been used with consistent and reliable results specially in the context of property prediction. Therefore, the first step in the development of a property prediction approach based on instance-based algorithms

was the definition, implementation and validation of a new structural similarity method, the Non-contiguous Atom Matching Structural Similarity function (NAMS), that is highly discriminative, comprehensive and at the same time able to verify the similarity principle. At the center of this methodology is the concept of atom alignment. This method is based on the comparison of bonding profiles of atoms on comparable molecules, including features that are seldom found in other structural or graph matching approaches like chirality or double bond stereoisomerism. The similarity measure is then defined on the annotated molecular graph, based on an iterative directed graph similarity procedure and optimal atom alignment using a pairwise matching algorithm. The atomic alignment approach often requires high computational cost, however, the similarities detected by the atom correspondence are more intuitively understood because similar atoms in the molecules are explicitly shown. Furthermore, considering that all similarity functions have a context that both define and limit their use, all defined atomic/bonds characteristics have a corresponding weight that can be adjusted or even eliminated in accordance with the context of the problem. New characteristics are also rather easy to include in the method. The number of parameters used by NAMS and their optimisation may seem a deterrent for its use, but the tests made suggest that despite the fact that individual similarity scores do change, the similarity patterns are identical when comparing large databases. Also, the empirical tests over three case-studies presented strongly suggest that predefined default parameter values able to provide coherent results is attainable. NAMS was compared with one of the most widely used similarity methods (Fingerprint-based similarity) for three case-studies with different objectives and characteristics. The method performed well and compared favourably to fingerprints for all 3 test cases. NAMS was able to distinguish either different or very similar hydrocarbons that were indistinguishable using a fingerprint-based approach and verifying the similarity principle using a dataset of very similar steroids with differences in the binding affinity to the corticosteroid binding globulin receptor. The method was also able to recover a significantly higher average fraction of active compounds when searching a database of highly diverse set of molecules with information about the MAO inhibition level. For this set it

was verified that the fraction of actives recovered per active seed searched, consistently increased with the similarity level, which further suggests that NAMS is actually capturing reliable structure-activity relationships. A good similarity method is useful to represent the compounds in a metric chemical space, however this is not enough to make good property/activity predictions and it can be complemented using interpolation techniques. As one of the objectives of this work was to design and implement tools to the community to make available the developed methodologies, a simple Web based tool is made available at http://nams.lasige.di.fc.ul.pt/. The full source code of the Python module is also freely available within the same website.

Taking into consideration that structurally similar molecules tend to have similar properties (Johnson & Maggiora, 1990), we developed a new method that takes into account the high dimensionality of the chemical space, predicting chemical, physical or biological properties using Kriging based on the most similar compounds in the molecular space composed by the instances of the training set and constructed based on their molecular similarity calculated by NAMS, consequently avoiding the selection of descriptors. Furthermore, the method takes into account the fact that similar molecules in the chemical space tend to yield similar property values while distant molecules can have very different values. The overall predictive results of this methodology applied to three different case-studies were good and within the confidence margins of the best models of other studies found in the literature. However the presented approach showed several advantages relatively to other QSPR/QSAR approaches, namely (1) it makes use of the dissimilarity/similarity between the compounds and it is not necessary to use any feature selection and no prior knowledge of the problem or property is necessary, thus it may be applicable to most QSAR/QSPR studies directly and to any possible compound (even to compounds that were never synthesized) as long as its structure is known, (2) the similarity map that positions each molecule of the dataset in the chemical space can be used to predict any chemical or biological property of the compounds as long as experimental data is available, (3) it is possible to identify the situations for which prediction errors are deemed to be higher (measure of extrapolation) by estimating the kriging variance for each prediction - high estimated kriging variances are obtained for compounds that are

out of the applicability of the model by the lack of similar neighbour compounds, while low estimated kriging variances indicate that the model is able to predict that property value with high confidence, (4) the model is readily understandable rather than a black box model, as it is possible to verify which of the compounds more directly impact the current estimation, (5) new compounds can be easily included as well as removed from the pool of training compounds since in this approach the target function is approximated locally for each test compound instead of an overall model that requires re-training each time the training set is changed, (6) the method can be applied to datasets of any size, however the predictive results are more likely to improve with the increase of the number of training instances as the probability of finding neighbour compounds with higher similarity increases, and (7) searches for the relationship among measured properties (richer hypothesis space) rather than approximate the modelled system by fitting the parameters of the selected basis functions (single hypothesis space). This approach can simultaneously solve multiple problems and deal successfully with changes in the problem domain.

The results of the application of this methodology also showed that the existence of at least one compound in the training set that has a high similarity with the test compound allows making predictions with higher confidence and reduced error. Another important conclusion is that the current approach can be used to guide the extension of the training set and exploration of new promising regions within the molecular space by suggesting new molecules that can be used as seed compounds for experimental property determination, which in turn will improve the model quality by providing a broader coverage of the molecular space, as well as being used for dataset curation proposes by analysing the prediction error and the structure/property of the selected neighbour compounds.

The estimated variance resulting from kriging for each prediction showed a strong correlation with the true prediction error, which proves that it can be used as a quality measure of each kriging prediction since it provides a confidence interval in the predicted value. It is our conviction that kriging estimated variance can also be used to interactively determine the number of compounds that should be used to make a prediction based on the minimization of the kriging estimated variance.

For all case-studies that were tested, the similarity function NAMS to map the compounds in the chemical space using random or temporal data selection yields better validated predictions with a smaller number of compounds (as nearest neighbours) for each prediction than using molecular descriptors or fingerprints. NAMS yielding better results was expected since a preliminary analysis comparing the pairwise distance between the compounds and their property difference showed that NAMS was able to discriminate the compounds better in accordance with the similarity principle: similar compounds tend to have similar property values and *vice-versa*. The application of a feature selection technique prior to the calculation of molecular similarity using molecular descriptors could be advantageous, however the objective of the study was to build a universal map of structural relationships in the chemical space that was not dependent on the property in study. In general, the predictive results are affected by redundancy between the compounds and by predictive maps that present patterns with several pairs of compounds at the same distance but with considerably different values of property and *vice-versa*. The primary output of this method is a function that maps structures to properties, i.e. given a structure drawn from the instance space it yields a prediction for its property value. Each prediction is based on the construction of a local approximation that applies in the neighbourhood of the new target instance.

The work presented herein is the first step in the creation of a new system for finding and developing promising compounds with desired target properties, resulting in the advancement of the current knowledge about the structure-property relationship. This framework (Figure 6.1) comprises several steps: (1) collect a set of molecular structures for which the target property is known (training set); (2) calculate the similarity between these structures using NAMS; (3) map these compounds in the metric space based on their structural similarity; (4) use a molecular structure generator (either to generate random molecules or to carry out small changes to molecules with known desired properties) or a set of test molecules; (5) calculate the similarity between these test and training molecules using NAMS; (6) interpolate their properties using kriging on a defined neighbourhood of the instance space; (7) repeat this process until a list of promising molecular structures that would be worthy to test in laboratory has been found.

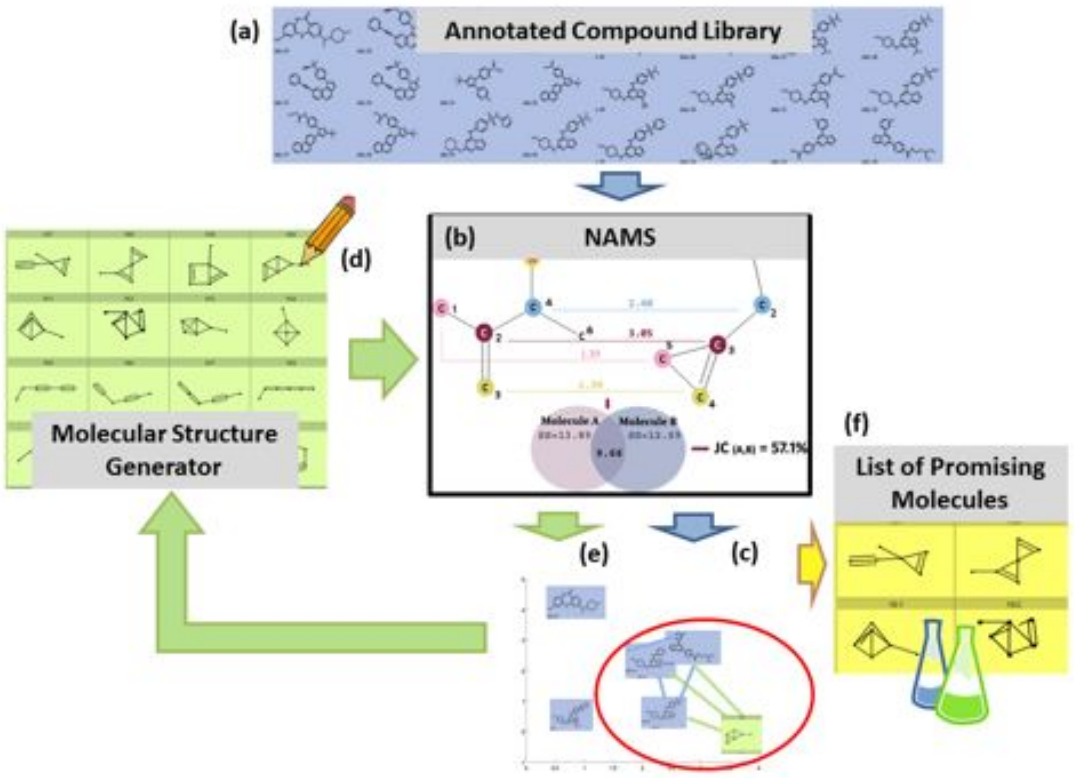


Figure 6.1: Framework for finding and developing promising compounds with desired target properties: a) Given a set of molecular structures for which properties are known (training set), b) calculate the similarity between these structures using NAMS and c) map these compounds in the metric space based on their structural similarity. d) Using a molecular structure generator or a set of test molecules, e) calculate the similarity between these test and training molecules using NAMS, interpolate their properties using kriging on a defined neighbourhood of the instance space and repeat this process until f) a list of promising molecular structures that would be worthy to test in laboratory has been found.

This framework offers a comprehensive approach to develop and find new compounds with desired properties as it is possible to clearly understand the effects of each structure change in the property value.

6.2 Summary of Research Contributions

The main contributions of this work were already described in detail thought this document, nevertheless these can be summarized as follows: **(1)** an innovative approach to improve the prediction power and comprehensibility of QSPR/QSAR problems using Random Forests for feature selection (Teixeira *et al.*, 2013b)); **(2)** the development of an Information System (ThermInfo) to collect, retrieve, and predict thermochemical data (Teixeira *et al.*, 2013c); **(3)** the development of an innovative structural similarity method (Noncontiguous Atom Matching Structural Similarity function (NAMS)) based on atom alignment between both molecules and a Web-tool that makes it available for the community (Teixeira & Falcao, 2013; Teixeira *et al.*, 2013a); **(4)** a new method to predict physical, chemical or biological properties of molecules using molecular structural similarity (Teixeira & Falcao, 2014); **(5)** collaboration in the development of new approaches based on Bayesian statistics coupled with machine learning and feature selection methods to produce robust models in real-world drug research scenarios (Martins *et al.*, 2012); **(6)** collaboration in the development of an Information System for Blood-Brain Barrier penetration data (B3Info).

6.3 Limitations and Future Work

Although the results presented here have demonstrated the effectiveness of the approaches, these could be further developed in a number of ways:

- The methodology that uses random forests, not as modelling tools for themselves, but as a method capable of identifying the most important features of a given modelling problem, which are then used as input variables to Support Vector Machines models is not specific for QSPR/QSAR problems. It would be interesting to test its performance in other predictive problems.

Although the algorithm was thoroughly described and it can be easily reproduced, its implementation as a R package would leverage its use by the community.

- The main bottleneck in the application of NAMS to calculate structural similarity between thousands of molecules is its computational cost which is mainly due to the nature of the method. Future improvements in its execution time may be achieved by rewriting the current implementation of NAMS in C programming language, parallelizing the calculations and using triangulation hierarchies (Jones & Ware, 1998) in order to implement neighbourhood search procedures. An implementation of NAMS is already being develop and it is able to calculate similarity between molecules in an average of 3 milliseconds which is about 19 times faster than the previous Python implementation.

- This study was limited to predict properties based solely on the structure of the compound as it may be used for any possible compound (even compounds that were never synthesized) and do not require any knowledge on their bioactivity or chemical/physical properties. Future work includes the development of a weighting schema in the similarity function to include both structural similarity and property profiles (using methods such as High Throughput Screening Fingerprints (Petrone *et al.*, 2012) or Similarity Ensemble Approaches (Keiser *et al.*, 2007) in order to accurately predict similar properties of compounds even when molecules are not structurally similar.

- Furthermore, although structurally-similar molecules are expected to exhibit similar properties, there are critical structural features that are able to drastically change biological properties of a compound leading to situations where a small structural change has a large impact in the property value or to the existence of compounds that are not structurally similar but similar in terms of properties (or *vice-versa*). For such situations hybrid approaches considering both general structural similarity and the existence of key-fragments or the compound role in biological processes could improve the predictive power of the models. Two possible ways of including such

information are: (1) a hybrid approach that integrates semantic similarity (Ferreira & Couto, 2010) based on ontologies for small molecules such as ChEBI (Degtyarenko *et al.*, 2008) with NAMS to calculate structural similarity; or (2) a hybrid approach that automatically identifies key substructural features in the compounds that are able to drastically change the property value and assigns them special weights in the structural similarity calculation, in order to guarantee that structurally similar compounds with small differences in the key substructure with have a highly penalized similarity score or *vice-versa*.

- Investigate and test more neighbourhood selection methods to optimize the predictive results of the Structural Similarity Based Kriging.
- Finally, the implementation of the framework (Figure 6.1) for finding and developing promising compounds with desired target properties in a user-friendly interface with different visualization tools of the process would leverage its use by the community and allow the eventual interest of the industry. This framework could also integrate property prediction visualization tools to facilitate the understanding of the obtained prediction, as well as functionalities to help in the curation and extension of datasets.

Appendix A

Case-Studies

The quality and performance of the methodologies developed in the context of this study were tested and validated in different real chemical or biological case-studies. This appendix is entirely dedicated towards the description of these case-studies (dependent variables of the model), collection and cleaning of datasets of experimental properties as well as the selection and preprocessing of different molecular representations (independent variables) which are assumed to influence the property of the molecule. The applicability and extrapolation capabilities of developed property prediction methodologies are highly conditioned by the size and quality of both datasets of properties and molecular representations.

A.1 Case-studies, Data and Data pre-processing

"If you torture the data enough, nature will always confess."
$\sim$ Ronald Coase (1981)

Data understanding is a crucial step that one should not overlook as it helps one to become familiar with the nature of the data prior to actual QSPR/QSAR model construction thereby reducing unnecessary errors or helping in the identification of interesting associations or relationships to study. However, before exploring data it is essential that thorough literature research on relevant background information pertaining to the biological or chemical system of interest is performed. Several data sets were applied to test and validate the methodologies

developed in this study. In the following subsections, these data sets are described with respect to their chemical/biological meaning, importance, origin, size and property range. A summary of all data sets is presented in Table A.1.

Data pre-processing can be considered to be one of the most important phases of property prediction modelling as it helps to ensure the integrity of the data set before proceeding further with data mining analysis. Essentially, the quality of a data mining analysis is a function of the quality of the data to be analysed. Therefore, for all case studies preliminary steps were performed to clean data in terms of anomalies, errors, or inconsistencies such as missing data, incomplete data, invalid character values, duplicated/redundant data, erroneous structures and consolidation of data collected from different sources which may cause problems when applying data mining algorithms.

A.1.1 Case A - Predicting Thermochemical Properties

Thermochemical properties, viz enthalpy of formation, enthalpy of vaporization, entropy, heat capacity are essential for the quantitative study of heat released or absorbed in a chemical reaction or transformation. Accordingly, thermochemical properties are particularly important in predicting the behaviour of reacting systems moving toward a more stable state of equilibrium. One of the most important state functions for a chemical system is the enthalpy, because it represents the ability to produce heat. For this reason, enthalpies of formation and of phase change were selected for this study. The enthalpy (H) comes from the Greek "to heat in" and is defined as the heat content of the system under constant pressure. It is not possible to directly measure the absolute value of enthalpy of a system. However, the enthalpy change (ΔH) of a chemical reaction can be measured and it represents the change in energy of a closed system due to chemical bonds being broken or formed and depends on the amount of reactants, the temperature and pressure of the system. The enthalpy change that occurs in a reaction is always calculated as the sum of the enthalpies of the products minus the sum of the enthalpies of the reactants. When heat is released, the change in the enthalpy for the system is negative (exothermic reaction), whereas when heat is absorbed, the change in the enthalpy is positive (endothermic reaction) (Atkins & Paula, 2001;

Table A.1: An overview of the case-studies.

Case-study	Property Type	Chemical/Biological Meaning	Task	N (train)	N (test)	Source
A1	Physico-Chemical	$\Delta_f H^\circ$ - gas phase for hydrocarbons	Regression	364	100	Teixeira *et al.* (2013c)
		$\Delta_f H^\circ$ - gas phase		1391	350	
		$\Delta_f H^\circ$ - liquid phase		1186	300	
A2		$\Delta_f H^\circ$ - crystalline phase		1159	300	
		$\Delta_{pc} H^\circ$ - fusion		63	20	
		$\Delta_{pc} H^\circ$ - vaporization		893	200	
		$\Delta_{pc} H^\circ$ - sublimation		464	150	
B		Aqueous solubility		1033	258/21	Huuskonen (2000)
C	Biological	Dihydrofolate Reductase Inhibition		237	124	Sutherland *et al.* (2004)
D		Population-level parameters of cytotoxicity		106	50	*
E		Steroids' binding affinity to the corticosteroid binding globulin receptor	Pattern Structure	31	-	Cramer *et al.* (1988)
F		Monoamine Oxidase Inhibition Level	Classification (4 classes)	1626	-	Brown & Martin (1996)
G		Blood-Brain Barrier Penetration	Classification (2 classes)	1850	120	Martins *et al.* (2012)

$\Delta_f H^\circ$ - Standard Molar Enthalpy of Formation
$\Delta_{pc} H^\circ$ - Standard Molar Enthalpy of Phase Change
N total number of instances * This dataset was provided by investigators at National Institute of Environmental Health Sciences (NIEHS), National Center for Advancing Translational Sciences (NCATS), and University of North Carolina (UNC), and were obtained through Synapse as part of the NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge (https://www.synapse.org/#!Synapse:syn1761567).

Masterton & Hurley, 2008). To insure a standardization of enthalpy measurements from different experimental sources and tabulation of data, a standard set of conditions are usually specified: one atmosphere (atm) pressure for the most stable form of a gaseous substance; one molar (M) concentration for substances in solution; one atm pressure for pure form of liquid and solid substances. The temperature is not part of the standard state definition but it should also be specified, normally 25°C (298.15 K) is used. Measurements made under these conditions are indicated by a superscript ° in the symbol of the quantity reported (Cox & Pilcher, 1970; Rossini, 1956; Stull *et al.*, 1969). There are two main factors affecting the standard enthalpy of formation: the relative strength of the bonds as measured by the bond enthalpies, and the relative number of bonds broken and formed (Masterton & Hurley, 2008). There are many practical applications of the study of these properties, namely solid and liquid fuel testing and classification, waste and refuse disposal, food and metabolic studies which extend into nutritional considerations and health concerns regarding the effects of diet on the human body, propellant and explosive testing and classification, industry efficiency and safety (e. g. Society for Testing and Materials (ASTM) regulations), and theoretical thermochemical studies stress different energy aspects such as bond energies, resonance energies and the nature of the chemical bond (Atkins & Paula, 2001; Fries, 1910; Sorai *et al.*, 2004; Stull *et al.*, 1969). Therefore, thermochemical data is an important resource for industrial research, however its availability is still scarce as this data is often expensive, difficult or in many cases impossible to measure. The experimental determination depends essentially on two components: (1) calorimetric part, which involves determination of the quantity of energy evolved or absorbed by the reaction or process and (2) chemical part, which involves measurements of the amount of the given reaction or process. The first part cannot be measured directly. The variation of energy during a reaction is measured indirectly by observing the change in temperature of a standard substance, which requires constant pressure conditions where virtually no heat is exchanged with the surroundings. The calorimetric part can be very difficult or impossible, since certain reactions are slow under normal conditions (for example the conversion of solid carbon from its graphite form to its

diamond form), other reactions will not happen until they are deliberately initiated (for example, some combustions must be ignited), other reactions involve a very small heat change, other reactions may evolve too much heat or form toxic substances which make them impossible to measure in the laboratory, etc. The second part requires examination of the purity of the chemical reaction and appropriate chemical or physical tests to demonstrate that the reaction which occurs in the calorimetric vessel is similar to the theoretically pure reaction or, if there is a side reaction, the amount and effect of it must be evaluated with the necessary accuracy (Atkins & Jones, 2007; Cox & Pilcher, 1970; Gislason & Craig, 2005; Rossini, 1956). In addition to all the difficulties faced in determining enthalpies of chemical compounds, experimental costs are never low, and they cannot be done on a shoestring. There are several costs that should be taken into account when making calorimetric experiments, such as equipment, reactants and human resources. Also due to difficulties to perform the experiment, it usually takes time and requires several repetitions to determine with some accuracy the enthalpy of a substance (Gislason & Craig, 2005; Rossini, 1956).

In this study two different subsets of a manually curated dataset (Teixeira *et al.*, 2013c) are used: *A1)* one for predicting enthalpy of formation in gas phase for hydrocarbons and *A2)* another one for predicting enthalpy of formation and phase change between different states for structurally diverse compounds. These datasets were also made available online at `http://therminfo.lasige.di.fc.ul.pt/`.

A.1.1.1 Case A1 - Predicting Enthalpy of formation for Hydrocarbon compounds

Hydrocarbon compounds are composed of carbon and hydrogen, sharing a key structural feature, the presence of stable carbon-carbon bonds. Although compounds of this class are structurally related, they exhibit important differences that are not easily distinguishable when using and comparing molecules in terms of the presence or absence of molecular features. For this reason and because hydrocarbon fragments are found in most types of compounds, a good prediction method should give an accurate and consistent estimation. Considering that the quality and prediction capabilities of any method strongly depend on the amount

and quality of the experimental data used for its development, the dataset used to model development was collected and manually curated by chemistry experts and it is available online on the ThermInfo database (Teixeira *et al.*, 2013c). The database contains 553 hydrocarbon compounds structurally characterized, but from these only 364 compounds have the experimental value for the standard molar enthalpy of formation in gas phase ($\Delta_f H_g^\circ$) at 298.15 K (Appendix B.1). The gas phase was selected due to its higher number of available experimental values. The distribution of the 364 compounds in the dataset into different types of hydrocarbons is presented in Table A.2.

Table A.2: Distribution of the compounds in the training and independent validation sets into the different types of hydrocarbons.

Type of hydrocarbon	**N (Train)**	**N (Test)**
Non-Cyclic	131	35
- Alkanes	66	7
- Alkenes	61	16
- Alkynes	4	12
Cyclic	233	65
- Aromatic	85	19
- Polycyclic	62	15
- Non-Aromatic	148	46
- Polycyclic	67	25
Hydrocarbons - Total	**364**	**100**

N - Total number of compounds

The $\Delta_f H_g^\circ$ values range from -705.8 kJ/mol for 11-Decylheneicosane to 780.9 kJ/mol for 5,6,11,12-Tetraphenylbenzo[b]anthracene, with a mean value of -33.6 kJ/mol and standard deviation of 190.8 kJ/mol. The distribution and variation of the dependent variable is shown in Figure A.1 - a). Although the values have a large range of distribution, the major part of the compounds' enthalpy is located between -500 and 500 kJ/mol.

The validation set was collected from two different sources, NIST Web book (version 2012) (Linstrom & Mallard, (accessed in 2012) and CRC Handbook of Chemistry and Physics (version 2010) (Lide, 2010). The validation set covers

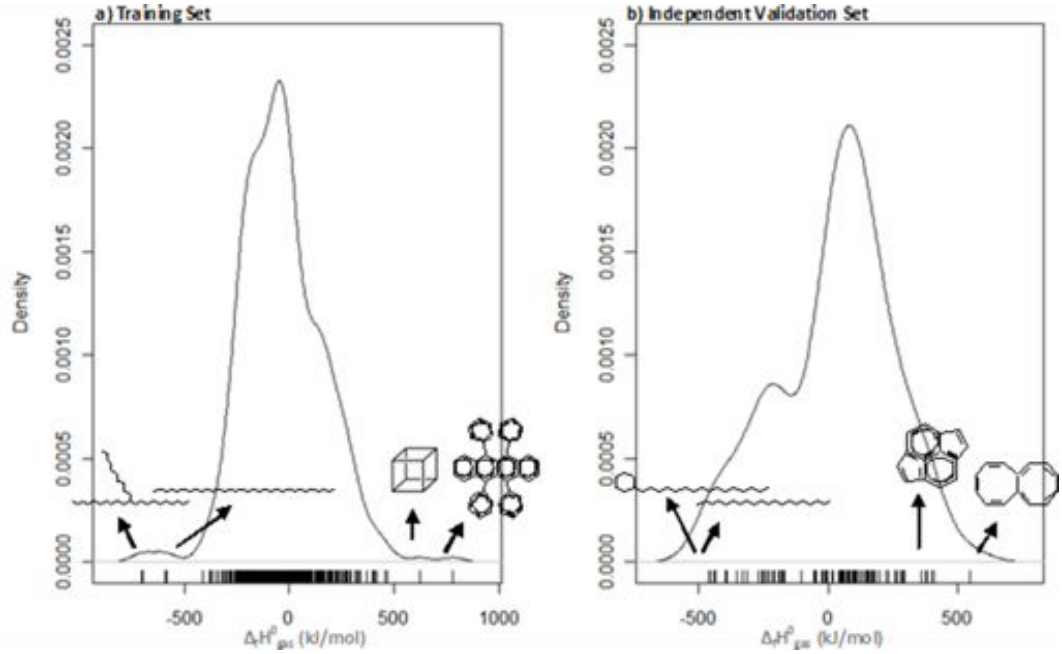


Figure A.1: Density plot showing the distribution and variation of the standard enthalpy of formation in gas phase at 298.15 K in the: **a)** training set, indicating the structure of the compounds with the most extreme values (maximum: 5,6,11,12-tetraphenylbenzo[b]anthracene and pentacyclo[4.2.0.0^{2,5}.0^{3,8}.0^{4,7}]octane; minimum: dotriacontane and 11-decylheneicosane); **b)** independent validation set, indicating the structure of the compounds with the most extreme values (maximum: (1Z,3Z,5Z,7Z,9Z,11Z)-octalene and cyclopenta[fg]acenaphthylene; minimum: hexadecylcyclohexane and eicosane).

different molecules that were not part of the training set and it contains 100 compounds randomly selected and structurally characterized and with experimental values for the $\Delta_f H_g^\circ$ at 298.15 K (Table A.1) (Appendix B.1). The $\Delta_f H_g^\circ$ values range from -460.50 kJ/mol to 551.50 kJ/mol, with a mean value of 30.02 kJ/-mol and standard deviation of 221.2 kJ/mol. The distribution and variation of the dependent variable is shown in Figure A.1 – b) and it is similar to the one obtained for the training set (Figure A.1 – a)).

A.1.1.2 Case A2 - Predicting Enthalpy of formation and phase change for ThermInfo's dataset

As mentioned before, the starting point of this case-study was to investigate property prediction in a smaller subset of structurally similar hydrocarbons. In this case-study we investigate the prediction of standard molar enthalpy of formation (gas, liquid and crystalline phases) and phase change (fusion, vaporization and sublimation) for a non-redundant set of highly diverse 2956 organic compounds. ThermInfo's dataset was manually curated by chemists in order to assure that it is thermodynamically consistent and experimental uncertainties of measurements are also included (Teixeira *et al.*, 2013c). All the compounds in the current version of the database are characterized with at least one thermochemical property. Table A.3 shows the number of compounds after a pre-processing step that are characterized with each thermochemical property in study and the distribution of these properties for the training and test sets (Appendix B.2).

Table A.3: Distribution of the compounds in the training and testing sets for six thermochemical properties in the ThermInfo dataset.

Thermochemical Property	**TRAIN**			**TEST**		
	N	**Property Range**	**Property Mean**	**N**	**Property Range**	**Property Mean**
Standard Molar Enthalpy of Formation ($\Delta_f H^\circ$) (kJ/mol)						
Gas phase	1391	[-4806.4, 780.9]	-205.1 ± 423.6	350	[-2564.4, 596.7]	-190.7 ± 348.4
Liquid phase	1186	[-4853.6, 814.3]	-297.4 ± 421.9	300	[-4019.2, 466.2]	-315.5 ± 472.7
Crystalline phase	1159	[-4499.0, 1218.6]	-312.0 ± 469.9	300	[-2236.7, 605.7]	-314.7 ± 410.3
Standard Molar Enthalpy of Phase Change ($\Delta_p H^\circ$) (kJ/mol)						
Fusion (solid to liquid)	63	[2.9, 72.0]	24.7 ± 14.3	20	[4.2, 58.2]	17.1 ± 10.9
Vaporization (liquid to gas)	893	[9.8, 142.3]	47.5 ± 16.5	200	[16.2, 122.6]	48.2 ± 16.8
Sublimation (solid to gas)	464	[31.9, 866.9]	98.8 ± 48.8	150	[43.0, 197.5]	95.4 ± 25.1

N - Total number of compounds

The distribution and variation of the Standard Molar Enthalpy of Formation (gas, liquid and crystalline) in the training and testing sets is shown in Figure A.2 – a) and b) and the Standard Molar Enthalpy of Phase Change (fusion, vaporization and sublimation) is shown in Figure A.2 – c) and d). In plots c) and d)

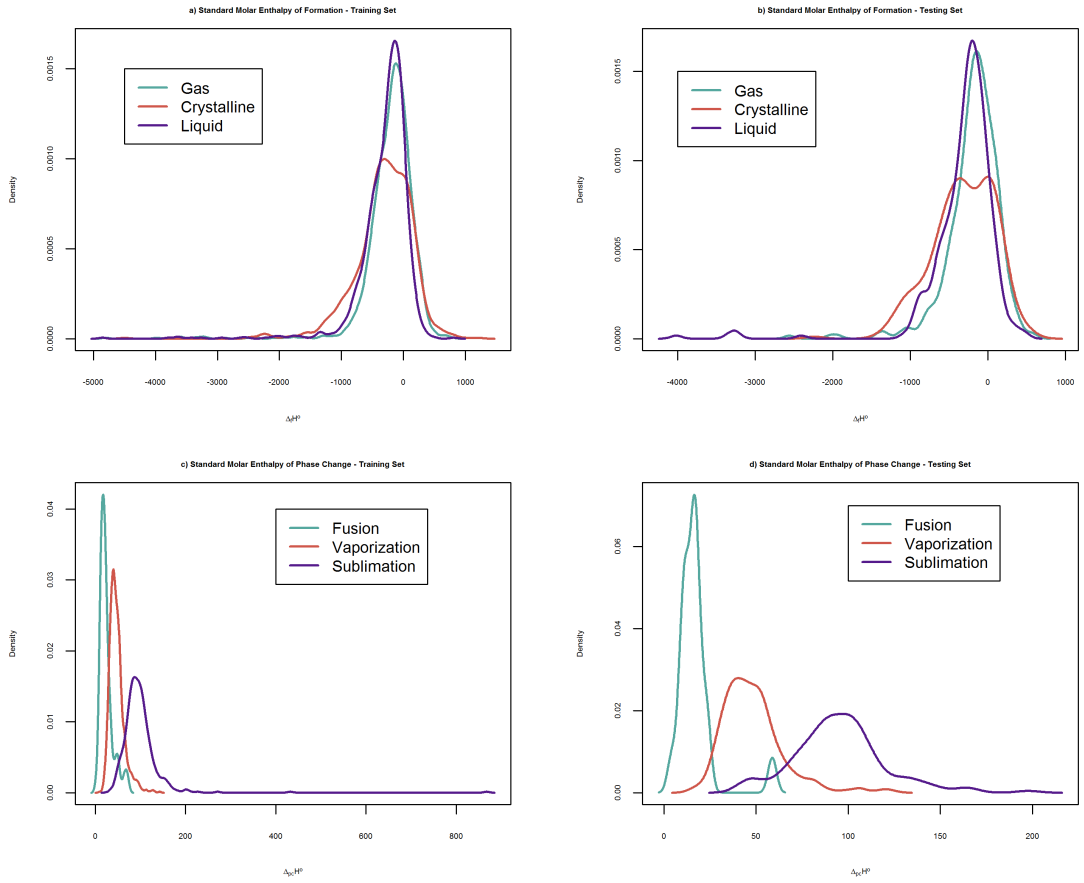


Figure A.2: Density plot showing the distribution and variation of **a)** Standard Molar Enthalpy of Formation in gas (1390 compounds), liquid (1186 compounds) and crystalline (1159 compounds) phases in kJ/mol for the ThermInfo's training set; **b)** Standard Molar Enthalpy of Formation in gas (350 compounds), liquid (300 compounds) and crystalline (300 compounds) phases in kJ/mol for the ThermInfo's testing set; **c)** Standard Molar Enthalpy of Phase Change of fusion (63 compounds), vaporization (893 compounds) and sublimation (464 compounds) phase transitions in kJ/mol for the ThermInfo's training set; and **d)** Standard Molar Enthalpy of Phase Change of fusion (20 compounds), vaporization (200 compounds) and sublimation (150 compounds) phase transitions in kJ/mol for the ThermInfo's testing set.

it is possible to verify that fusion, vaporization and sublimation are endothermic processes, since energy must be supplied to overcome the intermolecular forces that hold molecules together as a liquid or solid. Since the change from a solid

to a liquid, is not as big as the change from a liquid to a gas or a solid to a gas, enthalpies of fusion are smaller than enthalpies of vaporization or sublimation, respectively, for the same substance. The distribution and variation of the dependent variables is similar for the training and testing sets (Figure A.2).

A.1.2 Case B - Predicting Aqueous Solubility

Aqueous solubility is an important physical property of small organic molecules with pharmaceutical, environmental and industrial applications (Yalkowsky, 1999). It represents the maximum concentration of a chemical that will dissolve in pure water at a specified temperature. Aqueous solubility is a multi-mechanism system affected by different factors which makes predictions difficult because as more and structurally diverse compounds are measured, more data is obtained on more mechanisms, increasing the noise level (Lipinski, 2000; Salahinejad *et al.*, 2013). Several computational methods have been used to predict aqueous solubility using the structure of molecules, including group contribution methods (e.g. Gharagheizi *et al.* (2011); Klopman & Zhu (2001); Kuhne *et al.* (1995)), themodynamic calculations (e.g. Mirmehrabi *et al.* (2006); Palmer *et al.* (2008)), and quantitative structure-property relationships (e.g. Delaney (2004, 2005); Eric *et al.* (2012); Hughes *et al.* (2008); Huuskonen (2000); Salahinejad *et al.* (2013)).

The experimental aqueous solubility values used in this study for a total of 1291 diverse compounds were obtained from the literature (Cheminformatics.org, accessed in 2013; Huuskonen, 2000), which have been used for the development of several models (e.g. Hou *et al.* (2003); Huuskonen (2000); Liu & So (2001); Tetko *et al.* (2001); Yan & Gasteiger (2003)). The dataset was manually revised and several SMILES strings, compound names and CASRN were corrected, stereoisomeric information and electrical charges were incorporated in the SMILES strings and clusters composed of more than one fragment were simplified in order to maintain only the main molecule (dot-disconnected fragments such as ions were eliminated). The dataset of 1291 compounds was divided into a training set of 1033 compounds and a test set of 258 compounds (by selecting every fifth compound into the test set) as suggested by Liu & So (2001) (Appendix B.3). The model was also externally tested on a small set of 21 compounds that has been

extensively used for solubility prediction method validation since its introduction by Yalkowsky & Banerjee (1992). The aqueous solubility experimental measurements are reported as the negative logarithm of the molar solubility in water (mol/L) at temperatures between 20 and 25^{o}C. It is important to note that for the measurements obtained from relatively complex chemical structures, uncertainty in the experimental data should not be lower than approximately 0.5 log units (Huuskonen, 2000). The training set of 1033 compounds spans a solubility range of -11.62 to 1.58 log units with a mean value and standard deviation of -2.698 and 2.04 log units, respectively. The testing set of 258 compounds spans a solubility range of -9.15 to 1.13 log unit with a mean value and standard deviation of -2.83 and 2.02 log units, respectively. Finally, the external test set of 21 compounds spans a solubility range of -7.89 to -0.56 log units with a mean value and standard deviation of -3.83 and 1.86 log units, respectively. The distribution and variation of the dependent variable in the training and testing sets is shown in Figure A.3. It is possible to verify that the distribution of the properties in the training and testing 1 sets is similar, however the same does not verify for the testing set 2, increasing the difficulty in making predictions for this set.

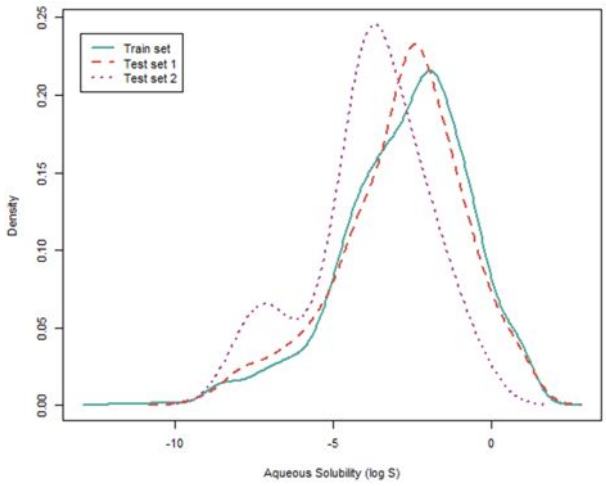


Figure A.3: Density plot showing the distribution and variation of aqueous solubility (log S) in the training (1033 compounds) and testing (test set 1: 258 compounds and test set 2: 21 compounds) sets.

A.1.3 Case C - Predicting Dihydrofolate Reductase (DHFR) Inhibition Activity

Dihydrofolate reductase (DHFR) is an enzyme that catalyses NADPH-dependent reduction of dihydrofolic acid to tetrahydrofolic acid, thus producing an important cofactor used in 1-carbon transfer reactions and is essential for the biosynthesis of purines, pyrimidines and amino acids (Banjanac *et al.*, 2009; Chen *et al.*, 1984). Inhibition of DHFR activity leads to a deficiency of thymidylate (dTMP), thus causing inhibition of cell growth (Chen *et al.*, 1984). They are used in the treatment and prophylaxis of major infectious diseases, such as malaria, toxoplasmosis and *Pneumocystis* pneumonia, as well as in the therapy of non-infectious human diseases such as psoriasis, inflammatory bowel disease, rheumatoid arthritis and neoplastic diseases (Banjanac *et al.*, 2009). Although structurally belonging to different classes, the majority of DHFR inhibitors contain 2,4-diamino substitution in pyrimidine ring (Banjanac *et al.*, 2009).

The dataset of DHFR inhibitors activity for rat liver has been taken from the study of Sutherland *et al.* (2004), in which it has been used to access the predictive accuracy of various methods encoding the molecular structure and using five different machine learning algorithms. The DHFR inhibition activity of a compound is measured in terms of half maximal inhibitory concentration (IC_{50}), which indicates the compound concentration that is needed to inhibit the DHFR by half. In this study, the IC_{50} values are converted to the pIC_{50} scale ($pIC_{50} = -log_{10}(IC_{50})$) in terms of molar concentration (mol/L). This dataset contains the pIC_{50} for 397 compounds which are divided in 237 compounds for the training set, 124 compounds for the test set and 36 inactive compounds with indeterminate activities ($IC_{50} > 10\mu M$) and that are used to verify if the models can correctly identify inactive compounds. It is important to note that the authors that compiled the dataset (Sutherland *et al.*, 2004) excluded all pairs of compounds with a Tanimoto similarity coefficient higher than 0.975 using for its calculation 2D structural fingerprints. The division of the compounds in training and test sets is the same as used in Sutherland *et al.* (2004) study that was based on the selection of 33% of the compounds to the test set by "cherry picking" with a maximum dissimilarity algorithm which maximizes the minimum squared

distance from each compound to all other compounds in the selected subset in order to maximize the diversity of the test set and to examine the predictive accuracy of the models when extrapolating outside the training set. The training set of 237 compounds spans a pIC_{50} range of 3.301 to 9.807 with a mean value and standard deviation of 6.226 and 1.268, respectively (Appendix B.4). The testing set of 124 compounds spans a pIC_{50} range of 3.569 to 9.398 with a mean value and standard deviation of 6.011 and 1.367, respectively (Appendix B.4). The distribution and variation of the dependent variable in the training and testing sets is shown in Figure A.4. It is possible to verify that the distribution of the properties in the training and testing sets is significantly different, increasing the difficulty in making predictions for this set.

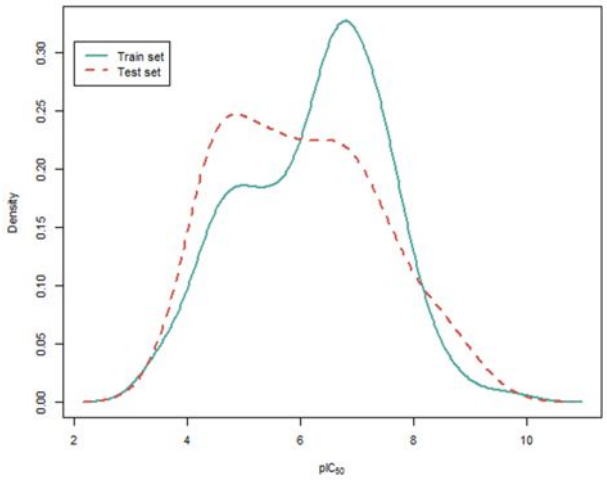


Figure A.4: Density plot showing the distribution and variation of DHFR inhibition activity (pIC_{50}) in the training (237 compounds) and testing (124 compounds) sets.

In this dataset each experimental measurement is associated to a reference from 1991 to 2002 making it possible to perform a temporal selection of training and test data. Thus, simulating a real-world scenario taking into account the appearance of new chemical series since earlier data will be used to predict later data. For that purpose, the dataset (including inactive compounds) was also divided based on the reference year of the property experimental measurement for each compound: 313 measurements obtained from 1991 to 1998 as training data to predict 84 measurements obtained from 1999 to 2002.

A.1.4 Case D - NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge

The core of this challenge (DREAM8 Consortium, 2013) was to build predictive models of cytotoxicity as mediated by exposure to environmental toxicants and drugs. To approach this question, a dataset containing cytotoxicity estimates as measured in lymphoblastoid cell lines derived from 884 individuals following in vitro exposure to 106 chemical compounds was provided (Figure A.5). The objective was to predict population-level parameters (median and interquartile distance) of cytotoxicity across 50 new chemicals based on their structural attributes. The computational models built within this challenge could be considered in certain decision-making contexts to inform government agencies as to which environmental chemicals and drugs are of the greatest potential concern to human health.

For that purpose, Sage Bionetworks, DREAM, University of North Carolina (UNC), the National Institutes of Environmental Health Sciences (NIEHS), and the National Center for Advancing Translational Sciences (NCATS) teamed up to generate a large population-scale toxicity screen in a human *in vitro* model system that leverages the 1000 Genomes Project. The lymphoblastoid cell lines are derived from 884 participants in the 1000 Genomes Project representing 9 distinct geographic sub-populations across Europe, Africa, Asia, and the Americas. Cell lines were selected to reflect unrelated individuals by removing all instances of first-degree relatives. These data are paired with the extensive, publicly available genomic data from these cell lines, including DNA variation profiles by the 1000 Genomes Project and transcriptomic data by the Geuvadis project. Cytotoxicity is represented in these data by the one-tenth maximal effective concentration (EC_{10}), which is a measure of potency calculated as the concentration of compound that induced a cytotoxic response equal to one-tenth of maximal cytotoxic response in that sample. A 10% "effect level" or "response" is a common point-of-departure in dose-response assessments of chemicals in human health. As such, models that quantitatively predict this adverse event (i.e., cytotoxicity) may have direct relevance to decisions on the potential of a chemical to cause harm at a certain dose. For each cell line, 156 unique chemical compounds were

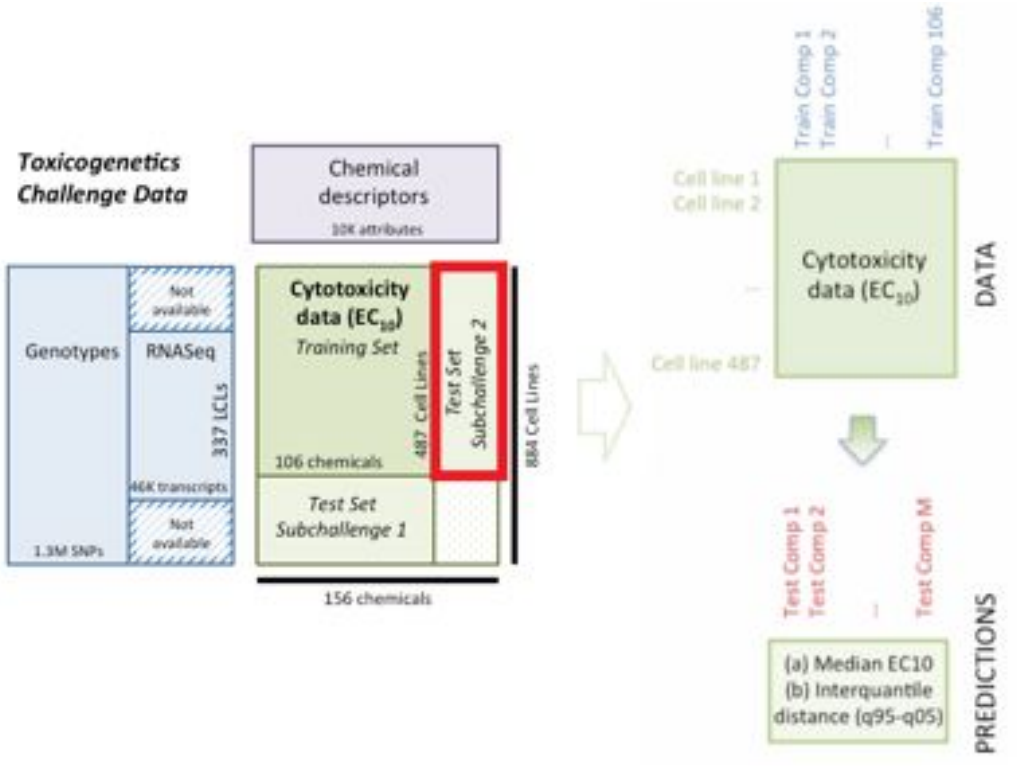


Figure A.5: General overview of the data available in NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge: **a)** training set including cytotoxicity estimates for each of 106 common environmental compounds (green); **b)** genomic profiles including SNP genotypes and RNAseq-based quantification of gene transcripts (blue); **c)** several thousand of chemical attributes based on chemical structure (purple). Highlighted in red is the sub-challenge described in this section, as well as a detailed workflow between data and prediction of population-level parameters of cytoxicity across chemicals. Adapted from: Webinar covering the NIEHS NCATS UNC DREAM Toxicogenetics Challenge: "Predicting toxicity from quantitative high-throughput screening in a population-based *in vitro* model (25th of July, 2013) by Lara Mangravite, Ivan Rusyn and Federica Eduati, available at http://www.youtube.com/watch?v=3JM2kv2gVjk

.

screened at 8 concentrations (0.33 nM to 92.2 μM) using a single 1536-well plate. Cytotoxicity was measured after forty hours at each dose by calculating intracellular ATP concentrations using the CellTiter-Glo Luminescent Cell Viability (Promega Corporation, Madison, WI) assay. EC_{10} values were calculated from concentration-response exposure data for each chemical across all 884 cell lines

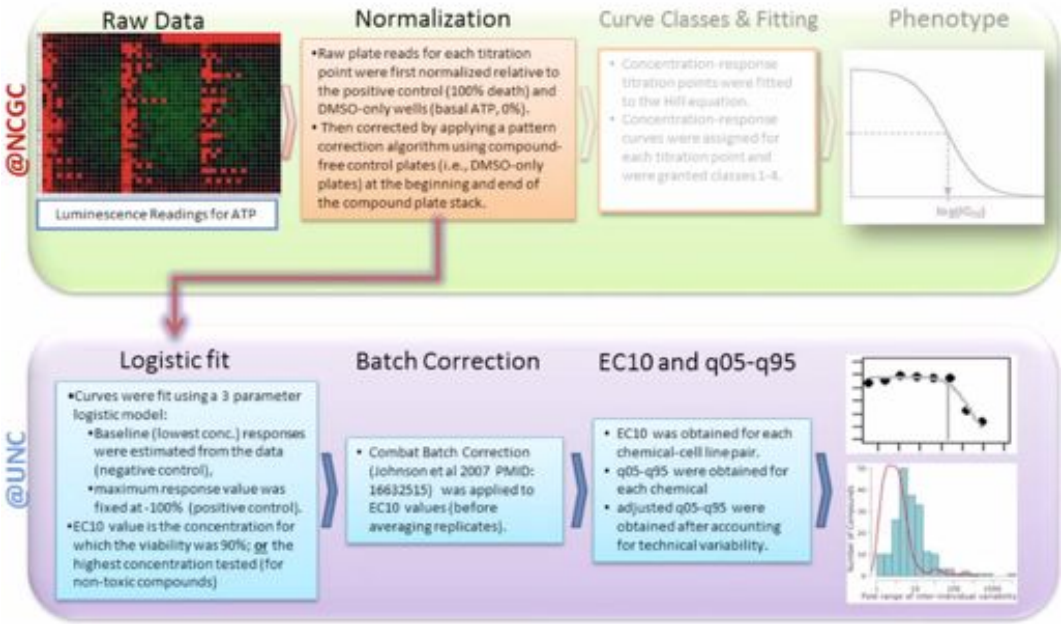


Figure A.6: Data Analysis and pre-processing scheme. Source: Webinar covering the NIEHS NCATS UNC DREAM Toxicogenetics Challenge: "Predicting toxicity from quantitative high-throughput screening in a population-based *in vitro* model (25th of July, 2013) by Lara Mangravite, Ivan Rusyn and Federica Eduati, available at http://www.youtube.com/watch?v=3JM2kv2gVjk

.

and were normalized relative to the positive/negative controls (Figure A.6). EC_{10} was defined as the concentration at which intracellular ATP content was decreased by 10% and was estimated for each cell line by normalizing data to vehicle treated cells and then fitting normalized concentration-response curves to a three parameter logistic regression model where maximum response was fixed to -100% and minimum response was derived from the response of the lowest three concentrations, with the exclusion of outliers as defined by >2 standard deviations. If the compound had less than 10% effect over the range of concentrations used in the experiment, the EC_{10} was set to 100 μM to represent a "no observable adverse effect level". Cell lines were then randomly divided into 5 screening batches with equal distribution of populations and gender in each batch. For each cell line, all chemical compounds were on the same 1536-well plate. Approximately 65% of the cell lines were seeded for repeat analysis on multiple plates (2-3 plates per

batch and/or between batches). EC_{10} values were batch corrected using Combat (Johnson *et al.*, 2007) and then replicate values per individual were averaged.

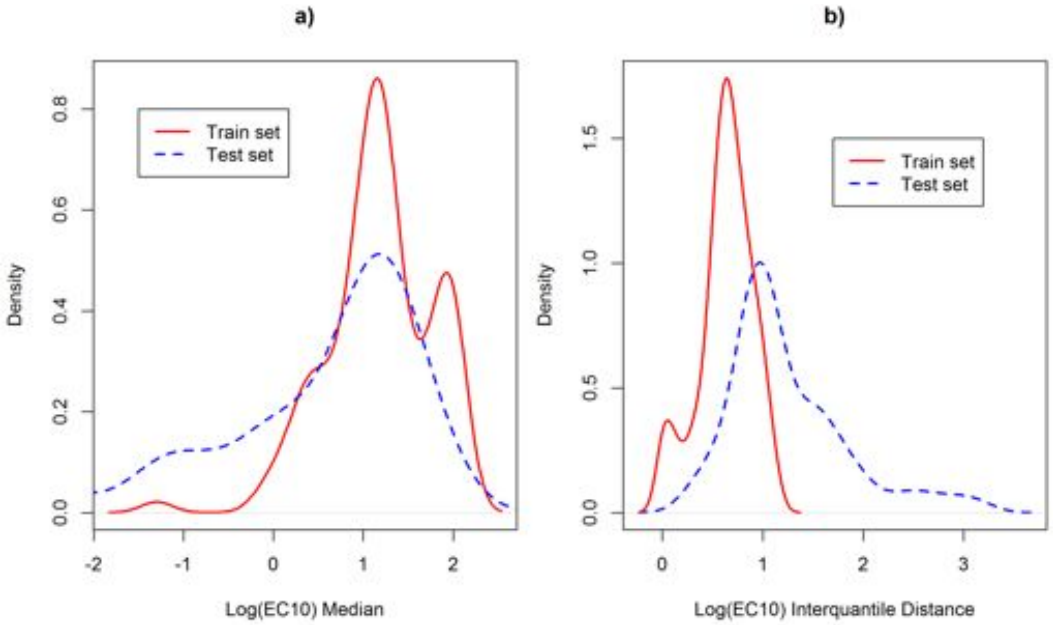


Figure A.7: Density plots showing the distribution and variation of **a)** logarithm of the median EC_{10} and **b)** logarithm of the interquantile distance in the training and testing sets of compounds.

The logarithm of the median EC_{10} for the training set varies from -1.30 to 2.00 with a mean value of 1.17 ± 0.59 while for the testing set varies from -2.55 to 1.95 with a mean value of 0.56 ± 1.06. The interquantile distance (95th - 5th quartiles) for the training set varies from 0.01 to 1.14 with a mean value of 0.63 ± 0.27 while for the testing set varies from 0.32 to 3.12 with a mean value of 1.25 ± 0.56. Figure A.7 displays the distribution and variation of the a) logarithm of the median EC_{10} and b) logarithm of the interquantile distance in the training and testing sets of compounds. It is easy to verify that the distribution of the properties in the training and testing sets are significantly different and biased, increasing the difficulty in making predictions for this set. The dataset and data terms of use are available at `https://www.synapse.org/#!Synapse:syn1761567`.

Potential utility of challenge outcomes are (1) understanding how chemical structure contributes to toxicity, (2) predicting toxicity for unknown or novel compounds from chemical structure information and (3) providing tools to help

predict potential human health hazard with the goal of informing regulatory decisions regarding existing and new compounds.

A.1.5 Case E - Steroids and their binding affinity to the corticosteroid binding globulin (CBG) receptor

Corticosteroid-binding globulin (CBG) is a monomeric glycoprotein with a single steroid-binding site which is a non-inhibitory member of the serine proteinase inhibitor (serpin) super-family that has an high-affinity to bind steroids in vertebrate blood (Lin *et al.*, 2010). Steroids are organic compounds that contain a characteristic arrangement (Figure A.8) composed of twenty carbon atoms bonded together in four fused rings (3 cyclohexane (A, B and C) and 1 cyclopentane (D) rings) and they vary mainly by the functional groups attached to four-ring core. Slight variations in this structure or in the atoms or groups attached to it produce profound differences in biological activity. Qualitatively, molecules with light substituents such as oxygen and hydroxyl at position 17 of steroid skeleton lead to low CBG activity, whereas the presence of the bulky chain such as $COCH_2OH$ enhances the activity. In contrast, the degree of aromaticity of the A ring is not important for biological activity.

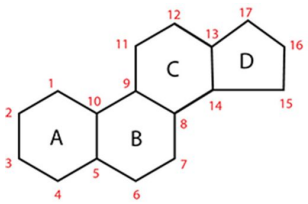


Figure A.8: The four-fused-ring steroid skeleton using letter designations for each ring and numbers for the carbon atoms.

The dataset comprises 31 steroids and their binding affinity (pK) to the CBG receptor in the human plasma compiled by Cramer *et al.* (1988) (Appendix B.5) and used in other studies regarding similarity calculations (e.g. Good *et al.* (1993)) and activity prediction (e.g. Anzali *et al.* (1996); Robert *et al.* (1999); Tuppurainen *et al.* (2002); Wagener *et al.* (1995)). The CBG binding is expressed by an affinity constant (K), which is converted to pK (equivalent to $-log(K)$).

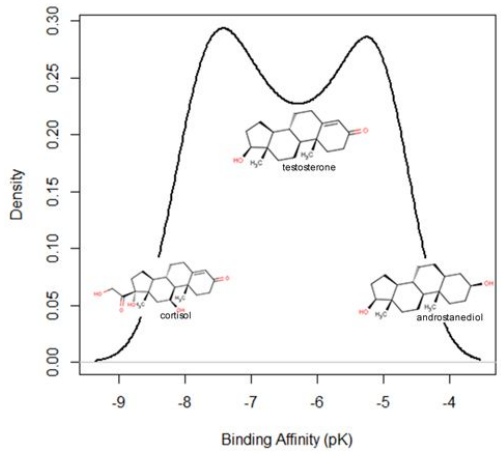


Figure A.9: Density plot showing the distribution and variation of CBG binding affinity (expressed as pK) of 31 steroids. Three representative compounds of the low (androstanediol), medium (testosterone) and high (cortisol) affinity levels are also depicted.

The more negative the pK value is, the higher the binding affinity. This dataset was chosen because although the structures have a similar structure, their activity level varies considerably from -5 to -7.881 with a mean value of -6.384 $\pm$ 1.082 and can be in 3 broad levels of affinity: low (> -6), medium (-7 to -6) and high (< -7). Considering that the binding strength of a receptor-substrate complex strongly depends on the shape of the substrate, the aim is to analyse the difference in the binding affinity of each pair of steroids to the CBG receptor solely based on the neighbourhood principle in order to characterize local aspects and discrimination of the data (pattern structure). Since biochemical activity of steroids varies considerably with seemingly small structural changes, this important molecular family presents a very challenging problem for any prediction method. Figure A.9 displays the distribution and variation of the binding affinity of the 31 steroids and three representative compounds of the low (androstanediol), medium (testosterone) and high (cortisol) affinity levels.

A.1.6 Case F - Classification of Monoamine Oxidase (MAO) Inhibition Level based on Molecular Similarity

Monoamine oxidase (MAO) is a flavin-containing enzyme tightly bound to the mitochondrial outer membrane of neuronal, glial, and other cells which catalyzes the oxidative deamination of several naturally occurring monoamines (Bach *et al.*, 1988). On the basis of their substrate and inhibitor specificities, two types of MAO (A and B) have been described: MAO-A preferentially deaminates serotonin, norepinephrine, and epinephrine and is irreversibly inhibited by low concentrations of clorgyline; MAO-B preferentially deaminates β-phenylethylamine and benzylamine and is irreversibly inhibited by deprenyl (Grimsby *et al.*, 1990). Due to their role in the metabolism of monoamine neurotransmitters, MAO-A and MAO-B present a considerable pharmacological interest. Inhibition of MAO-A exerts a antidepressant effect, while inhibition of both MAO-A and -B treats depression and anxiety. Since they enhance the dopaminergic tone in the brain, they are also used in the treatment of Parkinson's disease.

This dataset comprises 1650 considerably diverse compounds with information about MAO inhibition level (Accelrys, Inc., accessed in 2012; Brown & Martin, 1996). The activity is represented on a four-level scale: inactive compounds are represented as having activity 0, while the values 1, 2, and 3 correspond to increasing levels of activity. A pre-processing set was carried out where 5 molecules with unknown atoms were eliminated, 467 salts or clusters composed with more than one fragment were simplified in order to maintain only the main molecule (fragments such oxalic acid, sulfuric acid, sodium ion, among others were eliminated), molecules duplicated after fragment elimination (10 molecules) or with different affinity when clustered with other fragments (9 molecules) were eliminated. The final number of molecules is 1626 of which 288 are active (113 with activity level 1, 87 with activity level 2 and 88 with activity level 3) and 1338 are inactive. This dataset has been previously used to assess how structurally similarity methods based on Fingerprints relate with similar biological activity of molecules (Martin *et al.*, 2002). The dataset has an high number of actives, about 17.4% larger than what is typically found in screening databases, since it contains a subset of compounds synthesized to follow up a lead. Martin *et al.*

(2002) assessed the adequacy and bias of the dataset by determining the fraction of clusters of actives identified using different similarity thresholds (>0.85). Since the fraction of clusters containing active compounds increased with the increase of the threshold level, Martin *et al.* (2002) concluded that this dataset is not misleading and adequate for research. For this dataset, the aim is to retrieve compounds with similar activity level based on the similarity threshold using different metrics in order to assess their discriminative and predictive power.

A.1.7 Case G - Blood-Brain Barrier (BBB) Penetration Modelling

The Blood-Brain Barrier (BBB) is a membrane that separates circulating blood and the brain extracellular fluid. Some of the main functions of this barrier comprises the protection of the brain from substances in the blood that may injure it, protection against hormones and neurotransmitters in the rest of the body and maintenance of a constant environment for the brain(Pardridge, 1998). Therefore, the BBB has special features that make it almost impenetrable to most drugs. It has a selective permeability and the molecules that are generally able to cross the barrier have been difficult to identify. The features of the BBB represent a problem in CNS drug development, and most pharmaceutical companies do not have a BBB drug targeting development programme (King, 2011; Pardridge, 2005). BBB penetration is one of the key factors that are taken into account in chemical toxicological studies and in drug design (Zhang *et al.*, 2008). Furthermore, direct measurement of BBB penetration is possible but experiments are very expensive and time consuming (Zhao *et al.*, 2007) and constitute a time and financial hindrance when a large number of compounds are examined. Pardridge (2005) discusses the complexity of the process of BBB penetration, and how crucial its understanding is for treatment of several CNS disorders and even some viral infections like AIDS, where the virus lodges itself in brain tissues, where available antiviral drugs show minimal BBB penetration. Automated prediction of drug molecules' BBB penetration would be an useful tool to assist the experimental drug discovery process (Doniger *et al.*, 2000), decreasing the time of the initial stages and therefore the time required for a drug to reach the market.

A. CASE-STUDIES

For the present work, we compiled a dataset of 2053 molecules (Martins *et al.*, 2012) selected from a number of publications discussing BBB penetration (Doniger *et al.*, 2000; Li *et al.*, 2005; Zhang *et al.*, 2008; Zhao *et al.*, 2007). Of these, only 1970 were used for modelling purposes as all compounds that exceeded a molecular weight of 600 Da were excluded. Most studies divide molecules as being able or not to cross the BBB, that is belonging to BBB_+ or BBB_-, respectively. When the blood-brain penetration partition ($logBB$) is available, molecules were divided into BBB_+ and BBB_- classes if $logBB \geq -1$ and $logBB < -1$ respectively(Li *et al.*, 2005; Zhang *et al.*, 2008; Zhao *et al.*, 2007).

In total there are 1570 BBB_+ and 483 BBB_- (Appendix B.6). The dataset includes 324 molecules (179 BBB_+ and 145 BBB_-) from Doniger *et al.* (2000), 304 molecules (195 BBB_+ and 109 BBB_-) from Li *et al.* (2005), 100 molecules (91 BBB_+ and 9 BBB_-) from Zhang *et al.* (2008), and 1325 molecules (1105 BBB_+ and 220 BBB_-) from Zhao *et al.* (2007).

The resulting dataset was manually curated to remove generic name and SMILES duplicates and to treat some of the found ambiguities, e.g. Ribavirin, an anti-viral drug with broad spectrum, which it is ineffective against viral encephalitis because it fails to cross the BBB (Jeulina *et al.*, 2009), though in another study (Li *et al.*, 2005) the same molecule is described as capable of entering the CNS. The final dataset used for chemical descriptor generation includes a self-generated alphanumeric ID, the generic name as referred in the literature, the binary classification (p (BBB_+) or n (BBB_+)) and the respective SMILES representation (Appendix B.6). The structural molecular information was obtained using the Chemical Identifier Resolver (NCI/CADD CIR, 2011). Molecules where no recognizable name was found nor with a valid SMILES string, were eliminated. Furthermore molecules where contradicting data was found were also eliminated.

From the whole dataset of 1970 molecules (1455 BBB_+ and 395 BBB_-) a set of 120 (95 BBB_+ and 25 BBB_-) randomly selected molecules was withdrawn for constituting an independent validation set to be used in the final phase, after all the model selection procedures.

For this case-study the objective is to predict BBB penetration in real-world drug development scenarios, taking into account that this activity is a rare event

(only about 2% of the molecules are able to penetrate the BBB (Pardridge, 1998)), even though datasets in the literature are biased towards the number of BBB_+.

A.2 Molecular descriptors: implementation and pre-processing

"Ceci n'est pas une pipe."
∼ René Magritte (1929)

From a practical point of view, molecular descriptors are chemical information that is encoded within the molecular structure for which numerous methodologies are available for their calculation (Todeschini & Consonni, 2009; Todeschini *et al.*, 2008). Once molecular descriptors have been calculated they will serve as independent variables for further construction of QSPR/QSAR models, molecular similarity calculations or diversity analysis, assuming that there is an underlying relationship between molecular structure and properties. The key step in developing these models that reveal these relationship is the selection of an informative way to represent the molecules in the dataset. Several methodologies for the purpose of providing a standardized description of structural properties that are common across chemicals and can be used to model structure-based commonalities were studied, and here we present seven (descriptor sets **A-G**) that showed to be more relevant for the presented case-studies. These methodologies can be divided in three broad categories: (1) methods that represent the chemical compounds using quantitative descriptors of different nature (descriptor sets **A-C**); (2) methods that automatically analyse the structure of the chemical compound to identify a set of substructures that are relevant (descriptor sets **D-F**); (3) methods that map the compounds in the chemical space based on the similarity/dissimilarity between the structures (descriptor set **G**). Different combinations of these descriptors are also studied. One of the aspects that is considered important is the ability to interpret the models in a physico-chemical or biological sense. Thus, in the major part of the descriptor sets the selection of descriptors was limited to those that seem to carry some fundamental physico-chemical or biological information that might be related to the modelled property.

It is also important to note that there exists a great deal of variability in the range and distribution of these descriptors. This may pose a problem for machine learning algorithms as each descriptor will have different chance of contributing to the overall analysis. Such situation was handled by applying a data transformation step. The standard procedure consists of normalizing each variable to mean centring and variance scaling,

$$d'_i = \frac{(d_i - \bar{d}_i)}{s_{d_i}} \tag{A.1}$$

where d_i is the original vector of descriptors, $\bar{d}_i$ is the arithmetic mean of the vector of descriptors, s_{d_i} is the standard deviation and d'_i is the transformed vector of auto-scaled descriptors. This transformation is useful to ease interpretation and numerical stability, but it does not lead to changes in the coefficients, weights of the variables or in the interpretation of results. Furthermore, all zero variance descriptors (i.e. all the observations are the same) were removed.

A.2.1 Descriptor set A - Molecular descriptors calculated by E-DRAGON

E-DRAGON 1.0 (Tetko *et al.*, 2005; VCCLAB, accessed in 2011) is a free online version of DRAGON that was used to compute a set of 1666 molecular descriptors (2 and 3-Dimensional) commonly used in QSPR/QSAR analysis based on the compounds' SMILES. These include constitutional descriptors, walk and path counts, information indices, edge adjacency indices, topological charge indices, randic molecular profiles, radial distribution function descriptors, weighted holistic invariant molecular (WHIM) indices, functional group counts, charge descriptors, topological descriptors, connectivity indices, 2D autocorrelations, eigenvalue based indices, geometrical descriptors, 3D MoRSE descriptors, GETAWAY (GEometry, Topology, and Atom-Weights assemblY) descriptors, atom centred fragments as well as 31 other molecular properties (Todeschini & Consonni, 2009). The 3D atomic coordinates of the lower energy conformation for the provided molecules were calculated using CORINA (Sadowski *et al.*, 1994).

A.2.2 Descriptor set B - Structural descriptors calculated by Openbabel

The calculation of numerical specific molecular features was performed using the molecular SMILES by OpenBabel (O'Boyle *et al.*, 2011) and Pybel (O'Boyle *et al.*, 2008) libraries which in turn use *joelib2* (Guha *et al.*, 2006) for computation. These structural descriptors include, among others problem specific descriptors, molecular weight, average molecular weight, number of ring(s), number of bonds in ring(s) and atom multiplicity (number of primary, secondary, tertiary and quaternary carbon atoms). This set of descriptors was obtained in an objective way, selecting the descriptors that would represent the molecules in the dataset in a more appropriate way (e.g. distinguish structural isomers and compounds with multiple rings). Several combinations of descriptors were tried and selected based on preliminary predictive results.

A.2.3 Descriptor set C - Molecular descriptors calculated by Chemistry Development Kit (CDK)

The Chemistry Development Kit (CDK) is a freely available open-source Java library for Structural Chemo- and Bioinformatics (Steinbeck *et al.*, 2003). A set of 192 2D and 3D molecular descriptors was calculated using a Python binding to CDK. A detailed list of these descriptors can be found in the following reference: CDK (accessed in 2011).

A.2.4 Descriptor set D - Daylight Fingerprints

Daylight fingerprints (James *et al.*, 2011) have already been introduced in the precious chapter and they will be used as another option to represent molecules. Openbabel (O'Boyle *et al.*, 2011) was used to generate hashed bit-strings of 1024 bits (FP2) representing fragments up to seven atoms. The fingerprint computation requires the execution of the following steps: (1) obtain the fragments, (2) remove duplicate fragments and get a hash number for each fragment and (3) obtain the fingerprint of the fragments. The fragments are obtained by means of a recursive algorithm that finds all linear fragments up to a size of 7 atoms.

Cyclic fragments are also identified by checking if there are ring closures in the linear fragment. However, the algorithm does not identify branched fragments if they are not part of a ring. The hash number for a fragment is an integer number which is generated from the atoms and bonds of the fragment. The value of this number depends on the position, the atomic number and type (aromatic or not) of the atoms of the fragment, the position and the type of the bonds of the fragments, the position and the type of the ring closures. The fingerprint is a bit-string of 1024 bits although the three last bits of the fingerprint are always set to zero because 1021 is a prime number which produces a better hashing. It is evident that there are many more possible fragments than 1021, and consequently the number should be hashed dividing by 1021. The parameter set contains the binary fingerprint of each molecule with 1024 attributes, each was assumed to correspond to a distinct molecular descriptor: value 1 if the molecule has the corresponding fragment (bit is ON) and value 0 if the molecule does not have the corresponding fragment (bit is OFF).

A.2.5 Descriptor set E - Simplex representation of molecular structure (SiRMS)

Simplex representation of molecular structure (SiRMS) are molecular descriptors that represent counts of 9272 tetratomic fragments with fixed composition, structure, chirality and symmetry (Kuzmin *et al.*, 2005). The use of this descriptor set was limited to Case D - NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge, for which the organization generated a set of 9272 descriptors. Subsequently, another descriptor set, named binary SiRMS, was generated that instead of counts of fragments, the presence or absence of such fragments in the molecules were accounted for using 1 or 0, respectively.

A.2.6 Descriptor set F - Extended Laidler Bond Additivity (ELBA) parameters

Most physical and chemical properties are related to molecular structure namely atomic, bond, and group which contribute to the magnitude and type of inter-

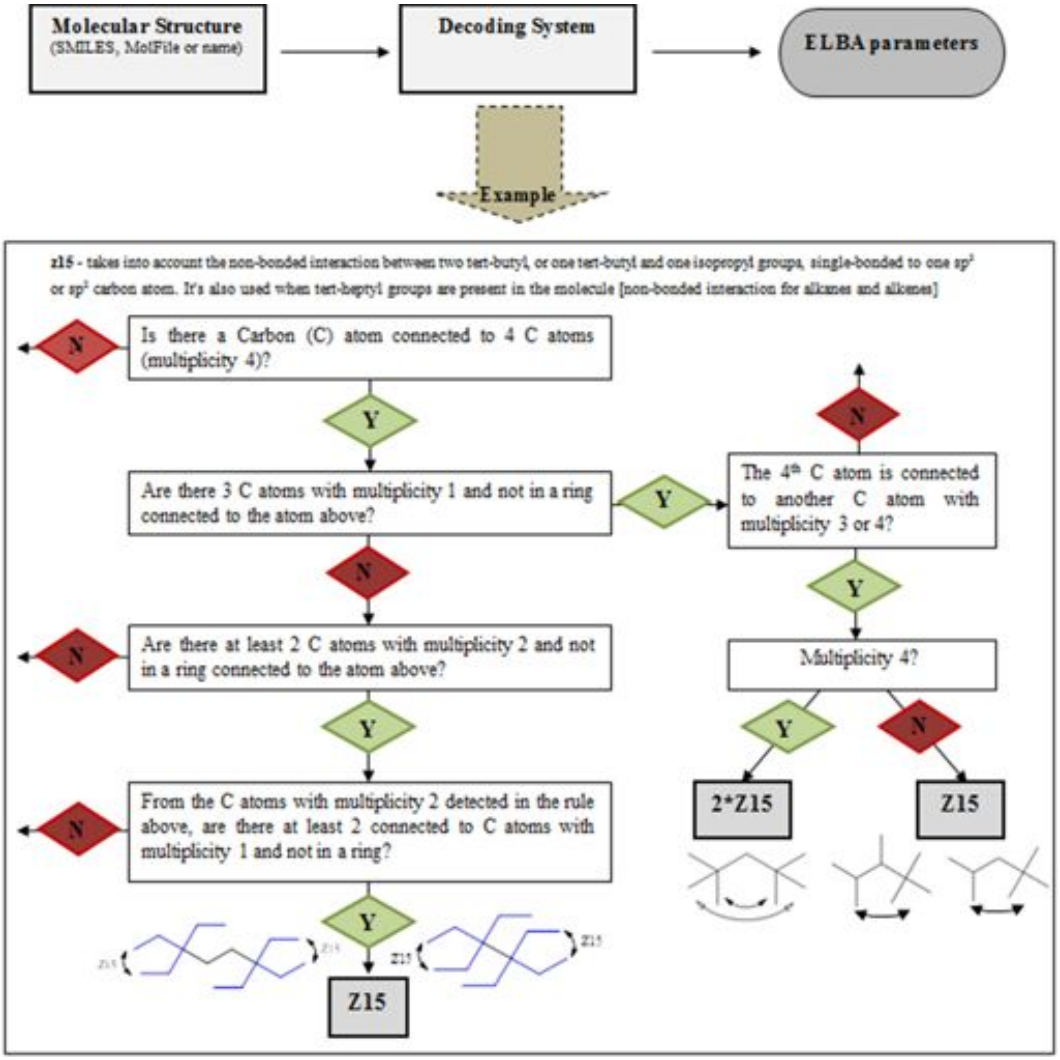


Figure A.10: Diagram representing the implementation of the ELBA parameters' generator with a specific example of the decoding system for the parameter z15.

molecular forces. This definition lead to the development of methods to estimate properties based on group/bond additivity (Benson & Buss, 1958; Laidler, 1956). These approaches divide the molecule into characteristic groups or bonds, then based on experimental measurements, a property contribution value is assigned to each group or bond, and the property value is obtained summing-up all of these values. A typical example of a refined bond additivity method is the Extended Laidler Bond Additivity (ELBA) method (Leal, 2006; Santos *et al.*, 2009). This additivity method includes a refined set of 165 parameters (Appendix B.7) that

account for additivity deviations (taking into account the chemical environment of the bond) in hydrocarbons. The advantage of this method is that each one of these parameters has an assigned physical meaning (i.e. they are not fudge parameters), whereas the disadvantage lies in its need for a high number of parameters to construct the method and for now it is restricted to hydrocarbon compounds. The automatic generation of these parameters were implemented in order to automatically iterate a computer-readable representation of the chemical structure to derive the frequencies of occurrence of the ELBA parameters. To extract the ELBA parameters a set of interrelated rules was implemented in Python based on identification of the presence or absence of certain structural characteristics using, for example, the number of atoms, number of bonds, number of single, double, triple and aromatic bonds, number of rings, atom multiplicity, bond order, connected atoms, maximum and minimum bond order, path-length, path-constitution, ring size and cis/trans bonds. The limitation of this descriptor set is that it can only be applied to hydrocarbon compounds. Figure A.10 shows in a schematic way the steps needed to calculate the descriptors using the implementation of the ELBA method and exemplifies the rules needed to extract the ELBA parameter *z15*.

A.2.7 Descriptor set G - Chemical Space Mapping based on Similarity/Dissimilarity

Molecular similarity or dissimilarity is defined through the intermolecular distance in the reference space. A meaningful chemical space represents a set of molecules and a set of associated relations (which can be similarities or dissimilarities). In this case two parameters are needed: a representation of the structure and a similarity/dissimilarity metric to map the compounds in the chemical space. Different structure representations have been applied in this study and are detailed in the following chapters. In chemistry it has generally been thought that, as most descriptor features are absent in most molecules, coefficients of association such as the Tanimoto are appropriate, since it only takes into account the proportion of "on-bits" that are shared by two compounds (Khalifa *et al.*, 2009).

Appendix B

Datasets and Implementation Details

B.1 Case-study A1: Datasets

- **Training Set.** [Available at: http://nams.lasige.di.fc.ul.pt/addfiles/a1_train.xlsx] Table containing information about the structure (ThermInfo ID, CASRN, compound name and SMILES) , the corresponding experimental values for the standard molar enthalpy of formation (kJ/mol) of gas phase at 298.15 K and the complete list of molecular descriptors for the compounds in the training set used in this study. More information about each compound can be found at http://therminfo.lasige.di.fc.ul.pt.

- **Independent Validation Set.** [Available at: http://nams.lasige.di.fc.ul.pt/addfiles/a1_test.xlsx] Table containing information about the structure (NIST Web book/CRC ID, CASRN, compound name and SMILES), the corresponding experimental values for the standard molar enthalpy of formation (kJ/mol) of gas phase at 298.15 K and the complete list of molecular descriptors for the compounds in the independent validation set used in this study. More information about each compound can be found in the CRC Handbook of Chemistry and Physics or NIST Chemistry WebBook (http://webbook.nist.gov/chemistry/).

- **NAMS subset of 100 Hydrocarbons.** [Available at: http://nams.lasige.di.fc.ul.pt/addfiles/a1_nams.xlsx] Table with compound name, CASRN, SMILES, similarity matrix using fingerprints, and similarity matrix using NAMS for the 100 hydrocarbons.

B.2 Case-study A2: Datasets

- **Training Set.** [Available at: http://nams.lasige.di.fc.ul.pt/addfiles/a2_train.xlsx] Table containing the ThermoInfo ID for each property in the training set. The availability of this dataset is restricted by data owner. Therefore only the IDs of the compounds are made available. All properties can be obtained searching ThermInfo's database by Molecular ID: http://therminfo.lasige.di.fc.ul.pt/search.php. The complete list of molecular descriptors calculated using CDK and Openbabel fingerprints for the compounds in the training set used in this study are also made available. More information about each compound can be found at http://therminfo.lasige.di.fc.ul.pt.

- **Test Set.** [Available at: http://nams.lasige.di.fc.ul.pt/addfiles/a2_test.xlsx] Table containing the ThermoInfo ID for each property in the test set. The availability of this dataset is restricted by data owner. Therefore only the IDs of the compounds are made available. All properties can be obtained searching ThermInfo's database by Molecular ID: http://therminfo.lasige.di.fc.ul.pt/search.php. The complete list of molecular descriptors calculated using CDK and Openbabel fingerprints for the compounds in the testing set used in this study are also made available. More information about each compound can be found at http://therminfo.lasige.di.fc.ul.pt.

B.3 Case-study B: Datasets

- **Data set - Aqueous Solubility.** [Available at: http://nams.lasige.di.fc.ul.pt/addfiles/b_data.xlsx] Table containing compound name,

SMILES, CASRN and aqueous solubility experimental values for training and testing sets.

B.4 Case-study C: Datasets

- **Dataset - Dihydrofolate reductase (DHFR) inhibitors activity.** [Available at: http://nams.lasige.di.fc.ul.pt/addfiles/c_data.xlsx] Table containing SMILES, experimental values for the DHFR inhibition activity in the rat liver, references, and publication year for the training and testing sets.

B.5 Case-study E: Datasets

- **Dataset - Steroids and their binding affinity to the corticosteroid binding globulin receptor.** [Available at: http://nams.lasige.di.fc.ul.pt/addfiles/e_data.xlsx] Table containing compound name, SMILES, binding affinity (pK), activity level, similarity matrix using fingerprints, and similarity matrix using NAMS for the 31 steroids (including the two atom substitution matrices (ASM = 0 and ASM = 1) used for the similarity calculation)

B.6 Case-study G: Datasets

- **Dataset - Blood-Brain Barrier (BBB) Penetration Modelling.** [Available at: http://nams.lasige.di.fc.ul.pt/addfiles/g_data.xlsx] Table containing compound name, SMILES and BBB penetration class for the training and testing sets.

B.7 Molecular descriptors F - Extended Laidler Bond Additivity (ELBA) parameters

- **Description of Extended Laidler Bond Additivity (ELBA) parameters.** [Available at: http://nams.lasige.di.fc.ul.pt/addfiles/elba_params.xlsx] Table containing all ELBA parameters and a short description as well as a link to a document containing a detailed description with examples.

B.8 NAMS - Atom Substitution Matrices (ASM)

- **Atom Substitution Matrices.** [Available at: http://nams.lasige.di.fc.ul.pt/addfiles/nams_asm.xlsx] File containing five different matrices: (1) ASM = 0, each atom type is only fully similar to itself (distance = 0) and completely different from all the others (distance = 1); ASM = 1, each atom type is only fully similar to itself (distance = 0) and partially different from all the others (distance = 0.9); ASM = 2 and 3, each atom type is only fully similar to itself (distance = 0) and partially different from all the others according to their position in terms of group and period in the periodic table (e.g. halogens are more similar); ASM = 4, all atoms are 100 % similar (distance = 0).

Appendix C

Results Details

C.1 Case-study A1: List of selected descriptors in model-based learning approach

- **List of descriptors selected using different selection/reduction methods: principal components analysis, genetic algorithms and variable importance calculated by random forests.** [Available at: http://nams.lasige.di.fc.ul.pt/addfiles/a1_descriptors.xls] For principal components analysis, the list of variables and respective factor loadings are presented for the ten first principal components (PC1 – PC10) , which are sufficient to explain 70.87% of the variance in the original dataset. For genetic algorithms, the number of times that each variable is selected in a total of 10 runs is presented. For variable importance calculated by random forests, a list of the variables is presented, along with their average and standard deviation of the importance score in the ten runs (ordered according to the average variable importance score).

C.2 Case-study A2: Detailed results for model-based learning approach

- **Detailed results for case-study A2.** [Available at: http://nams.lasige.di.fc.ul.pt/addfiles/results_a2.xls] Detailed table of results obtained

for all models in the training and testing sets using different descriptor sets for all properties of case-study A2.

- **Detailed results for testing set of case-study A2.** [Available at: http://nams.lasige.di.fc.ul.pt/addfiles/results_test_a2.xls] Detailed table of predictive results obtained for all properties of case-study A2 using the best model (selected based on training cross-validated results).

- **List of descriptors selected for each property of case-study A2.** [Available at: http://nams.lasige.di.fc.ul.pt/addfiles/a2_listdesc.xls] For variable importance calculated by random forests, a list of the variables is presented, along with their average and standard deviation of the importance score in the ten runs (ordered according to the average variable importance score) for each property in this case-study.

C.3 Case-study G: Detailed results for model-based learning approach

- **Detailed results for case-study G using a combination of descriptor sets A, B and D.** [Available at: http://nams.lasige.di.fc.ul.pt/addfiles/g_abd.xlsx] Detailed Table presenting variable importance of all molecular descriptors as well as results obtained using descriptor sets A, B and D in the training set of case-study G to model BBB penetration.

- **Detailed results for case-study G using a combination of descriptor sets B and D.** [Available at: http://nams.lasige.di.fc.ul.pt/addfiles/g_bd.xlsx] Detailed Table presenting variable importance of all molecular descriptors as well as results obtained using descriptor sets B and D in the training set of case-study G to model BBB penetration.

C.4 Case-study A1: Detailed results for neighbourhood selection in instance-based learning approach

- **Detailed results for neighbourhood selection using case-study A1.** [Available at: http://nams.lasige.di.fc.ul.pt/addfiles/neighs_results.xlsx] Detailed Table presenting predictive results of structural-based kriging using NAMS, molecular descriptors and fingerprints, different neighbourhood selection methods and parametrizations. It is also shown detailed results for computation time required for prediction using the best predictive configuration.

C.5 Case-studies B and C: Detailed predictive results using an instance-based learning approach based on Kriging

- **Detailed results for neighbourhood selection using case-study A1.** [Available at: http://nams.lasige.di.fc.ul.pt/addfiles/inst_c_b_results.xlsx] Detailed table with the predictive results obtained using *DistKrig* and *CoordKrig* coupled with Molecular Descriptors, Fingerprints, and NAMS and with a different number of compounds selected for training.

References

Abraham, A. (2005). Artificial neural networks. In P.H. Sydenham & R. Thorn, eds., *Handbook of Measuring System Design*, chap. 3, 901–908, John Wiley & Sons, Ltd. 52

Aha, D.W. & Kibler, D. (1991). *Instance-based learning algorithms*, 37–66. Kluwer Academic Publishers, Oxford. 153

Accelrys, Inc. (accessed in 2012). Mao dataset, pipeline pilot v8.5. San Diego CA. USA. 236

Andrew, R.L. (2001). *Molecular modelling: principles and applications*. Prentice Hall. 58

Anzali, S., Barnickel, G., Krug, M., Sadowski, J., Wagener, M., Gasteiger, J. & Polanski, J. (1996). The comparison of geometric and electronic properties of molecular surfaces by neural networks: Application to the analysis of corticosteroid-binding globulin activity of steroids. *J. Comput.-Aided Mol. Des.*, **10**, 521–534. 234

Arlot, S. & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statist. Surv.*, **4**, 40–79. 58

Ash, S., Cline, M.A., Homer, R.W., Hurst, T. & Smith, G.B. (1997). Sybyl line notation (sln): A versatile language for chemical structure representation. *J. Chem. Inf. Comput. Sci.*, **37**, 71–79, doi: 10.1021/ci960109j. 28

Atkins, P. (2003). *Atkins' Molecules*. Cambridge University Press. 3

REFERENCES

Atkins, P. & Jones, L. (2007). *Chemical Principles: The Quest for Insight*. W H Freeman & Co. 221

Atkins, P. & Paula, J.d. (2001). *Physical Chemistry*. W. H. Freeman, New York, 7th edn. 218, 220

Auer, J. & Bajorath, J. (2008). Molecular similarity concepts and search calculations. In J.M. Keith, ed., *Bioinformatics*, vol. 453 of *Methods in Molecular Biology*, chap. 17, 327–347, Humana Press, Totowa, NJ. 114

Bach, A.W., Lan, N.C., Johnson, D.L., Abell, C.W., Bembenek, M.E., Kwan, S.W., Seeburg, P.H. & Shih, J.C. (1988). cdna cloning of human liver monoamine oxidase a and b: molecular basis of differences in enzymatic properties. *Proc. Natl. Acad. Sci. U.S.A.*, **85**, 4934–4938. 236

Bachrach, S. (2009). Chemistry publication - making the revolution. *J. Cheminf.*, **1**, Article 2, http://www.jcheminf.com/content/1/1/2 (accessed September, 2013). 1, 3

Bajorath, J. (2001). Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.*, **41**, 233–245, doi: 10.1021/ci0001482. 114

Bajorath, J. (2004). *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*. Biomed Protocols, Humana Press, Totowa, NJ. 1, 2, 25, 36, 41, 59, 119

Bajorath, J. & Warr, W.A. (2011). Some trends in chem(o)informatics. In J. Bajorath, ed., *Chemoinformatics and Computational Chemical Biology*, vol. 672 of *Methods in Molecular Biology*, 1–37, Humana Press. 2

Balakin, K. (2009). *Pharmaceutical Data Mining: Approaches and Applications for Drug Discovery*. Hoboken, NJ, USA: John Wiley & Sons. 44, 45, 46, 64

Baldi, P. & Brunak, S. (2001). *Bioinformatics - the machine learning approach (2. ed.)*. MIT Press. 57

Banjanac, M., Tatic, I., Ivezic, Z., Tomic, S. & Dumic, J. (2009). Pyrimido-pyrimidines: a novel class of dihydrofolate reductase inhibitors. *Food Technol. Biotech.*, **47**, 236–245. 228

Barnard, J.M. (1993). Substructure searching methods: Old and new. *J. Chem. Inf. Comput. Sci.*, **33**, 532–538, doi: 10.1021/ci00014a001. 118

Barnard, J.M., Jochum, C.J. & Welford, S.M. (1989). *A Universal Structure/Substructure Representation for PC-Host Communication*, vol. 400 of *ACS Symposium Series*, 76–81. American Chemical Society, doi:10.1021/bk-1989-0400.ch008. 28

Basak, S.C., Gute, B.D., Mills, D. & Hawkins, D.M. (2003). Quantitative molecular similarity methods in the property/toxicity estimation of chemicals: a comparison of arbitrary versus tailored similarity spaces. *J. Mol. Struct.*, **622**, 127–145. 155

Batista, J., Godden, J.W. & Bajorath, J. (2006). Assessment of molecular similarity from the analysis of randomly generated structural fragment populations. *J. Chem. Inf. Model.*, **46**, 1937–1944, doi: 10.1021/ci0601261. 118

Baumann, K., Darvas, F. & Schneider, G. (2008). Qsar & combinatorial science: Transition to the future. *QSAR Comb. Sci.*, **27**, 5–5. 23

Bayram, E., Santago, P., Harris, R., Xiao, Y.D., Clauset, A. & Schmitt, J. (2004). Genetic algorithms and self-organizing maps: a powerful combination for modeling complex qsar and qspr problems. *J. Comput.-Aided Mol. Des.*, **18**, 483–493. 48

Bender, A. & Glen, R.C. (2004). Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem*, **2**, 3204 – 3218. 40, 41, 114, 115, 116, 134, 167

Bender, A., Mussa, H.Y., Glen, R.C. & Reiling, S. (2003). Molecular similarity searching using atom environments, information-based feature selection, and a naive bayesian classifier. *J. Chem. Inf. Comput. Sci.*, **44**, 170–178, doi: 10.1021/ci034207y. 40

Bender, A., Jenkins, J.L., Scheiber, J., Sukuru, S.C.K., Glick, M. & Davies, J.W. (2009). How similar are similarity searching methods? a principal component analysis of molecular descriptor space. *J. Chem. Inf. Model.*, **49**, 108–119, doi: 10.1021/ci800249s. 115

Benson, S. & Buss, J. (1958). Additivity rules for the estimation of molecular properties thermodynamic properties. *J. Chem. Phys.*, **29**, 546–572. 243

Berglund, A.E. & Head, R.D. (2010). Pzim: A method for similarity searching using atom environments and 2d alignment. *J. Chem. Inf. Model.*, **50**, 1790–1795, doi: 10.1021/ci1002075. 118

Biau, G. (2012). Analysis of a random forests model. *J. Mach. Learn. Res.*, **98888**, 1063–1095. 71

Blum, A. & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artif. Intell.*, **97**, 245–271. 64

Bohling, G. (2005). Introduction to geostatisticsa and variogram analysis. [Online] http://people.ku.edu/~gbohling/cpe940/Variograms.pdf (accessed November, 2013). 163

Breiman, L. (2001). Random forests. *Mach. Learn.*, **45**, 5–32. 53, 54, 65, 71

Bronstein, A., Bronstein, M. & Kimmel, R. (2008). *Numerical geometry of non-rigid shapes*. Springer, New York. 119

Broughton, M. & Queener, S. (1991). Pneumocystis carinii dihydrofolate reductase used to screen potential antipneumocystis drugs. *Antimicrob. Agents Chemother.*, **35**, 1348–55. 188

Brown, A.C. & Fraser, T.R. (1868). On the connection between chemical constitution and physiological action. part. i. on the physiological action of the salts of the ammonium bases, derived from strychnia, brucia, thebaia, codeia, morphia, and nicotia. *Earth Env. Sci. T. R. So.*, **25**, 151–203. 20

Brown, F.K. & James, A.B. (1998). *Chemoinformatics: What is it and How does it Impact Drug Discovery*, vol. 33, book section 35, 375–384. Academic Press. 2

Brown, R.D. & Martin, Y.C. (1996). Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.*, **36**, 572–584. 134, 219, 236

Brown, R.D. & Martin, Y.C. (1997). The information content of 2d and 3d structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.*, **37**, 1–9. 134

Brown, R.D., Jones, G., Willett, P. & Glen, R.C. (1994). Matching two-dimensional chemical graphs using genetic algorithms. *J. Chem. Inf. Comput. Sci.*, **34**, 63–70, doi: 10.1021/ci00017a008. 118

Bunin, B.A. (2007). *Chemoinformatics: theory, practice, & products*. Springer. 44

Burbidge, R., Trotter, M., Buxton, B. & Holden, S. (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis. *BMC Neurol.*, **26**, 5–14. 52

Burden, F.R. (2001). Quantitative structure-activity relationship studies using gaussian processes. *J. Chem. Inf. Model.*, **41**, 830–835. 156

Burges, C.J. (1998). A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, **2**, 121–167. 53, 70

Cahn, R.S., Ingold, C. & Prelog, V. (1966). Specification of molecular chirality. *Angew. Chem. Int. Ed. (English)*, **5**, 385–415. 136, 138

Chambers, J.M. (1992). Linear models. In T.J. Hastie, ed., *Statistical Models in S*, Wadsworth and Brooks/Cole, Pacific Grove, California. 167

Chen, M.J., Shimada, T., Moulton, A.D., Cline, A., Humphries, R.K., Maizel, J. & Nienhuis, A.W. (1984). The functional human dihydrofolate reductase gene. *J. Biol. Chem.*, **259**, 3933–43. 228

Chen, W.L. (2006). Chemoinformatics: Past, present, and future. *J. Chem. Inf. Model.*, **46**, 2230–2255. 1, 2, 63

Chih-Chung, C. & Chih-Jen, L. (2001). Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 1–27. 70

Cho, S. & Hermsmeier, M. (2002). Genetic algorithm guided selection: Variable selection and subset selection. *J. Chem. Inf. Comput. Sci.*, **42**, 927–936. 65

CDK (accessed in 2011). Cdk descriptor summary. http://pele.farmbio.uu.se/nightly-1.2.x/dnames.html. 241

Cheminformatics.org (accessed in 2013). Qsar datasets - huuskonen data set. http://cheminformatics.org/datasets/huuskonen/index.html. 226

CRAN-R (2012). Principal components analysis. http://stat.ethz.ch/R-manual/R-patched/library/stats/html/prcomp.html. 45

Consonni, V., Ballabio, D. & Todeschini, R. (2009). Comments on the definition of the q2 parameter for qsar validation. *J. Chem. Inf. Model.*, **49**, 1669–1678, pMID: 19527034. 58

Cooley, W. & Lohnes, P. (1971). *Multivariate Data Analysis*. New York: J. Wiley and Sons Inc. 45, 46, 47

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Mach. Learn.*, **20**, 273–297. 66, 70

Cover, T. & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory.*, **13**, 21–27. 153

Cox, J.D. & Pilcher, G. (1970). *Thermochemistry of organic and organometallic compounds*. Academic Press, London and New York. 220, 221

Cramer, R. (1993). Partial least squares (pls): Its strengths and limitations. *Perspect. Drug Discov.*, **1**, 269–278. 46

Cramer, R.D., Patterson, D.E. & Bunce, J.D. (1988). Comparative molecular field analysis (comfa). 1. effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.*, **110**, 5959–5967. 219, 234

Darvas, F., Papp, A., Bagyi, I., Ambrus, G. & Urge, L. (2004). Openmolgrid, a grid based system for solving large-scale drug design problems. In M. Dikaiakos, ed., *Grid Computing*, vol. 3165 of *Lecture Notes in Computer Science*, 69–76, Springer Berlin Heidelberg. 24

Davis, J. (2002). *Statistics and Data Analysis in Geology*. John Wiley and Sons, New York. 156

Dearden, J.C., Cronin, M.T.D. & Kaiser, K.L.E. (2009). How not to develop a quantitative structure-activity or structure-property relationship (qsar/qspr). *SAR QSAR Environ. Res.*, **20**, 241–266. 63

Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., Mcnaught, A., Alcántara, R., Darsow, M., Guedj, M. & Ashburner, M. (2008). Chebi: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344–D350. 215

Dehmer, M., Varmuza, K., Bonchev, D. & Emmert-Streib, F. (2012). *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*. Weinheim, Germany: Wiley-VCH Verlag GmbH. 64, 65

Delaney, J. (1996). Assessing the ability of chemical similarity measures to discriminate between active and inactive compounds. *Mol. Divers.*, **1**, 217–222. 134

Delaney, J.S. (2004). Esol: Estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.*, **44**, 1000–1005. 226

Delaney, J.S. (2005). Predicting aqueous solubility from structure. *Drug Discov. Today*, **10**, 289–295. 226

Deutsch, C.V. (1996). Correcting for negative weights in ordinary kriging. *Computers & Geosciences*, **22**, 765–773. 165

Diaz-Uriarte, R. & Alvarez de Andres, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3. 66

Dickson, M. & Gagnon, J.P. (2009). The cost of new drug discovery and development. *Discovery Medicine*, **4**, 172–179. 5

Diggle, P.J. & Jr, P.J.R. (2007). *Model Based Geostatistics*. Springer, New York. 166

Dixon, S.L. & Merz, K.M. (2001). One-dimensional molecular representations and similarity calculations: Methodology and validation. *J. Med. Chem.*, **44**, 3795–3809, doi: 10.1021/jm010137f. 34

DREAM8 Consortium (2013). Niehs-ncats-unc dream toxicogenetics challenge. https://www.synapse.org/#!Synapse:syn1761567. 230

Doniger, S., Hofmann, T. & Yeh, J. (2000). Predicting cns permeability of drug molecules: Comparison of neural network and support vector machine algorithms. *J. Comput. Biol.*, **9**, 849–864. 237, 238

Doucet, J.P. & Panaye, A. (2011). *QSARs in Data Mining*, 253–266. QSAR in Environmental and Health Sciences, CRC Press, Boca Raton. 63

Du Xihua, C.Y. & Keying, C. (2009). The qspr research on the environmental contaminants of chlorinated aromatic compounds. *Comp. Appl. Chem.*, **1**, 013. 21

Dudek, A.Z., Arodz, T. & Galvez, J. (2006). Computational methods in developing quantitative structure-activity relationships (qsar): a review. *Comb. Chem. High Throughput Screen*, **9**, 213–228. 20, 44, 46, 47, 49, 51, 52, 53, 70

Dutta, D., Guha, R., Wild, D. & Chen, T. (2007). Ensemble feature selection: Consistent descriptor subsets for multiple qsar models. *J. Chem. Inf. Model.*, **47**, 989–997. 63

Eckert, H. & Bajorath, J. (2007). Molecular similarity analysis in virtual screening: foundations limitations and novel approaches. *Drug Discov. Today*, **12**, 225–233. 41, 114, 116

Ehrlich, H.C. & Rarey, M. (2011). Maximum common subgraph isomorphism algorithms and their applications in molecular science: a review. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **1**, 68–79. 117, 118, 169

Ekins, S., Mestres, J. & Testa, B. (2007). In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *Br. J. Pharmacol.*, **152**, 9–20. 21

Eklund, M., Norinder, U., Boyer, S. & Carlsson, L. (2014). Choosing feature selection and learning algorithms in qsar. *J. Chem. Inf. Model.*, **54**, 837–843. 51, 52

Eric, S., Kalinic, M., Popovic, A., Zloh, M. & Kuzmanovski, I. (2012). Prediction of aqueous solubility of drug-like molecules using a novel algorithm for automatic adjustment of relative importance of descriptors implemented in counter-propagation artificial neural networks. *Int. J. Pharm.*, **437**, 232–241. 226

Eriksson, L., Andersson, P., Johansson, E. & Tysklind, M. (2006). Megavariate analysis of environmental qsar data. part i: A basic framework founded on principal component analysis (pca), partial least squares (pls), and statistical molecular design (smd). *Mol. Divers.*, **10**, 169–186. 45, 46

Falcao, A.O., Langlois, T. & Wichert, A. (2006). Flexible kernels for rbf networks. *Neurocomputing*, **69**, 2356–2359, doi: 10.1016/j.neucom.2006.03.006. 61

Fang, K.T., Yin, H. & Liang, Y.Z. (2004). New approach by kriging models to problems in qsar. *J. Chem. Inf. Comput. Sci.*, **44**, 2106–2113. 156

Faulon, J.L. & Bender, A. (2009). *Handbook of Chemoinformatics Algorithms*. Taylor and Francis. 33, 39, 59

Ferreira, J.D. & Couto, F.M. (2010). Semantic similarity for automatic classification of chemical compounds. *PLoS Comput. Biol.*, **6**, e1000937. 215

Flower, D.R. (1998). On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.*, **38**, 379–386, doi: 10.1021/ci970437z. 38, 39, 116, 133, 134, 155, 168

Fodor, I. (2002). A survey of dimension reduction techniques. Tech. rep., Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.8.5098. 45, 46, 47, 49

Fries, J.A. (1910). *Methods and Standards in Bomb Calorimetry*. HardPress. 220

Frohlich, H., Wegner, J. & Zell, A. (2004). Towards optimal descriptor subset selection with support vector machines in classification and regression. *QSAR Comb. Sci.*, **23**, 311–318. 65

Froidevaux, R. (1993). Constrained kriging as an estimator of local distribution functions. In *Proceedings of the International Workshop on Statistics of Spatial Processes: Theory and Applications. Bari, Italy*, 106–118. 165

Gakh, A.A., Burnett, M.N., Trepalin, S.V. & Yarkov, A.V. (2011). Modular chemical descriptor language (mcdl): Stereochemical modules. *J. Cheminf.*, **3**, 5. 136

Garcia, G.C., Ruiz, I.L. & Gomez-Nieto, M.A. (2003). Step-by-step calculation of all maximum common substructures through a constraint satisfaction based algorithm. *J. Chem. Inf. Comput. Sci.*, **44**, 30–41, doi: 10.1021/ci034167y. 118

Garey, R. & Johnson, D.S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, New York. 118, 121

Garrett, D., Peterson, D., Anderson, C. & Thaut, M. (2003). Comparison of linear, nonlinear, and feature selection methods for eeg signal classification. *IEEE Trans. Neural. Syst. Rehabil. Eng.*, **11**, 141–144. 48

Gasteiger, J. (2003). *Handbook of Chemoinformatics: From Data to Knowledge*. Weinheim: Wiley-VCH. 1, 23, 25, 29, 31, 32, 33, 34, 36, 38, 39, 63

Genuer, R., Poggi, J.M. & Tuleau, C. (2008). Random forests: some methodological insights. Tech. rep., INRIA Université Paris XI - Paris Sud, http://hal.inria.fr/inria-00340725/en/. 65, 71, 72

Genuer, R., Poggi, J.M. & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern. Recognit. Lett.*, **31**, 2225–2236. 65, 66, 71, 72

Geronikaki, A., Druzhilovsky, D., Zakharov, A. & Poroikov, V. (2008). Computer-aided prediction for medicinal chemistry via the internet. *SAR QSAR Environ. Res.*, **19**, 27–38, doi: 10.1080/10629360701843649. 34

Gharagheizi, F., Eslamimanesh, A., Mohammadi, A.H. & Richon, D. (2011). Representation/prediction of solubilities of pure compounds in water using artificial neural network - group contribution method. *J. Chem. Eng. Data*, **56**, 720–726. 226

Ghosh, P. & Bagchi, M. (2009). Qsar modeling for quinoxaline derivatives using genetic algorithm and simulated annealing based feature selection. *Curr. Med. Chem.*, **16**, 4032–4048. 49

Gillet, V.J., Wild, D.J., Willett, P. & Bradshaw, J. (1998). Similarity and dissimilarity methods for processing chemical structure databases. *Comput. J.*, **41**, 547–558, 10.1093/comjnl/41.8.547. 119

Gislason, E.A. & Craig, N.C. (2005). Cementing the foundations of thermodynamics: Comparison of system-based and surroundings-based definitions of work and heat. *J. Chem. Thermodyn.*, **37**, 954–966, doi: 10.1016/j.jct.2004.12.012. 221

Golbraikh, A. & Tropsha, A. (2002). Beware of q2! *J. Mol. Graphics Modell.*, **20**, 269–276, doi: 10.1016/S1093-3263(01)00123-1. 55, 58, 59

Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Boston, MA: Addison-Wesley Professional. 47

Goldberg, D. & Holland, J. (1988). Genetic algorithms and machine learning. *Mach. Learn.*, **3**, 95–99. 47, 48

Gonzalez, M.P., Teran, C., Saiz-Urra, L. & Teijeira, M. (2008). Variable selection methods in qsar: An overview. *Curr. Top. Med. Chem.*, **8**, 1606–1627. 64

Good, A.C., So, S.S. & Richards, W.G. (1993). Structure-activity relationships from molecular similarity matrices. *J. Med. Chem.*, **36**, 433–438. 234

Goodarzi, M., Heyden, Y.V. & Funar-Timofei, S. (2013). Towards better understanding of feature-selection or reduction techniques for quantitative structure-activity relationship models. *Trends Anal. Chem.*, **42**, 49–63. 46

Gramatica, P. (2007). Principles of qsar models validation: internal and external. *QSAR Comb. Sci.*, **26**, 694–701. 45, 55, 58, 67, 82, 87

Gramatica, P., Vighi, M., Consolaro, F., Todeschini, R., Finizio, A. & Faust, M. (2001). Qsar approach for the selection of congeneric compounds with a similar toxicological mode of action. *Chemosphere*, **42**, 873 – 883, chemistry for Protection of the Environment. 46

Gramatica, P., Giani, E. & Papa, E. (2007). Statistical external validation and consensus modeling: A qspr case study for koc prediction. *J. Mol. Graphics Modell.*, **25**, 755 – 766. 58

Grimsby, J., Lan, N.C., Neve, R., Chen, K. & Shih, J.C. (1990). Tissue distribution of human monoamine oxidase a and b mrna. *J. Neurochem.*, **55**, 1166–1169. 236

Guha, R. & Jurs, P.C. (2004). Development of linear, ensemble, and nonlinear models for the prediction and interpretation of the biological activity of a set of pdgfr inhibitors. *J. Chem. Inf. Comput. Sci.*, **44**, 2179–2189, pMID: 15554688. 47

Guha, R., Howard, M.T., Hutchison, G.R., Murray-Rust, P., Rzepa, H., Steinbeck, C., Wegner, J.K. & Willighagen, E.L. (2006). The blue obelisk-interoperability in chemical informatics. *J. Chem. Inf. Model.*, **46**, 991–998. 29, 241

Gunda, T.E. (2011). Chemical drawing programs - the comparison of isis and symyx draw, chemdraw, drawit (chemwindow), acd/chemsketch and chemistry 4-d draw. Electronic article, University of Debrecen, Hungary. 29

Gute, B.D. & Basak, S.C. (2001). Molecular similarity-based estimation of properties: a comparison of three structure spaces. *J. Mol. Graphics Modell.*, **20**, 95–109. 155

Hagadone, T.R. (1992). Molecular substructure similarity searching: efficient retrieval in two-dimensional structure databases. *J. Chem. Inf. Comput. Sci.*, **32**, 515–521, doi: 10.1021/ci00009a019. 118

Hammett, L.P. (1935). Some relations between reaction rates and equilibrium constants. *Chem. Rev.*, **17**, 125–136. 21

Han, J., Kamber, M. & Pei, J. (2011). *Data Mining: Concepts and Techniques*. San Francisco, USA: Morgan Kaufmann, 3rd edn. 45, 61

Hand, D.J., Mannila, H. & Smyth, P. (2001). *Principles of data mining*. MIT Press. 61

Hansch, C. & Fujita, T. (1964). p-σ-π analysis. a method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.*, **86**, 1616–1626. 21

Hawe, G.I., Alkorta, I. & Popelier, P.L.A. (2010). Prediction of the basicities of pyridines in the gas phase and in aqueous solution. *J. Chem. Inf. Model.*, **50**, 87–96. 156

Heikamp, K. & Bajorath, J. (2011). Large-scale similarity search profiling of chembl compound data sets. *J. Chem. Inf. Model.*, **51**, 1831–1839, doi: 10.1021/ci200199u. 134

Henikoff, S. & Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 10915–10919. 135

Holliday, J.D., Hu, C.Y. & Willett, P. (2002). Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2d fragment bit-strings. *Comb. Chem. High Throughput Screen*, **5**, 155–166. 42, 43, 116, 119

Hou, T.J., Xia, K., Zhang, W. & Xu, X.J. (2003). Adme evaluation in drug discovery. 4. prediction of aqueous solubility based on atom contribution approach. *J. Chem. Inf. Comput. Sci.*, **44**, 266–275. 177, 226

Hu, G., Kuang, G., Xiao, W., Li, W., Liu, G. & Tang, Y. (2012). Performance evaluation of 2d fingerprint and 3d shape similarity methods in virtual screening. *J. Chem. Inf. Model.*, **52**, 1103–1113, doi: 10.1021/ci300030u. 115

Hughes, L.D., Palmer, D.S., Nigsch, F. & Mitchell, J.B.O. (2008). Why are some properties more difficult to predict than others? a study of qspr models of solubility, melting point, and log p. *J. Chem. Inf. Model.*, **48**, 220–232. 226

Huuskonen, J. (2000). Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.*, **40**, 773–777. 177, 219, 226, 227

Isaaks, E. & Srivastava, R. (1989). *An Introduction to Applied Geostatistics*. Oxford University Press, New York. 154, 156, 159

Islam, M., Mahdi, J. & Bowen, I. (1997). Pharmacological importance of stereochemical resolution of enantiomeric drugs. *Drug Safety*, **17**, 149–165. 120

Itskowitz, P. & Tropsha, A. (2005). k nearest neighbors qsar modeling as a variational problem: theory and applications. *J. Chem. Inf. Model.*, **45**, 777–785. 50

James, C., Weininger, D. & Delany, J. (2011). *Daylight Theory Manual - version 4.9*. Laguna Niguel, CA: Daylight Chemical Information Systems, Inc. 28, 36, 38, 241

Jaworska, J., Nikolova-Jeliazkova, N. & Aldenberg, N. (2005). Qsar applicabilty domain estimation by projection of the training set descriptor space: a review. *ATLA, Altern. Lab. Anim.*, **5**, 445–459. 23

Jeulina, H., Venard, V., Carapito, D., Finance, C. & Kedzierewicz, F. (2009). Effective ribavirin concentration in mice brain using cyclodextrin as a drug carrier: Evaluation in a measles encephalitis model. *Antiviral Res.*, **81**, 261–266. 238

Johnson, M.A. & Maggiora, G.M. (1990). *Concepts and Applications of Molecular Similarity*. Wiley, New York. 5, 7, 19, 20, 114, 141, 154, 155, 173, 179, 209

Johnson, W.E., Li, C. & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, **8**, 118–127. 233

Jones, C.B. & Ware, J.M. (1998). Proximity search with a triangulated spatial model. *Comput. J.*, **41**, 71–83. 151, 214

Journel, A.G. & Rao, S.E. (1996). Deriving conditional distributions from ordinary kriging. *Stanford Center for Reservoir Forcasting, Stanford University Report*, 25. 165

Karatzoglou, A., Meyer, D. & Hornik, K. (2006). Support vector machines in r. *J. Stat. Softw.*, **15**, 1–28. 70

Karelson, M. (2000). *Molecular descriptors in QSAR/QSPR*. New York: John Wiley & Sons. 63

Katritzky, A., Lobanov, V. & Karelson, M. (1995). Qspr: the correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.*, **24**, 279–287. 63

Katritzky, A., Karelson, M. & Lobanov, V. (1997). Qspr as a means of predicting and understanding chemical and physical properties in terms of structure. *Pure App. Chem.*, **69**, 245–248. 63

Katritzky, A.R. & Fara, D.C. (2005). How chemical structure determines physical, chemical, and technological properties: An overview illustrating the potential of quantitative structure-property relationships for fuels science. *Energy and Fuels*, **19**, 922–935. 21

Katritzky, A.R., Maran, U., Lobanov, V. & Karelson, M. (2000). Structurally diverse quantitative structure-property relationship correlations of technologically relevant physical properties. *J. Chem. Inf. Comput. Sci.*, **40**, 1–18. 21, 63

Katritzky, A.R., Petrukhin, R., Tatham, D., Basak, S., Benfenati, E., Karelson, M. & Maran, U. (2001). Interpretation of quantitative structure-property and -activity relationships. *J. Chem. Inf. Comput. Sci.*, **41**, 679–685, pMID: 11410046. 45

Katritzky, A.R., Fara, D.C., Petrukhin, R.O., Tatham, D.B., Maran, U., Lomaka, A. & Karelson, M. (2002). The present utility and future potential for medicinal chemistry of qsar / qspr with whole molecule descriptors. *Curr. Top. Med. Chem.*, **24**, 1333–1356. 63

Kawabata, T. (2011). Build-up algorithm for atomic correspondence between chemical structures. *J. Chem. Inf. Model.*, **51**, 1775–1787, doi: 10.1021/ci2001023. 118

Keiser, M.J., Roth, B.L., Armbruster, B.N., Ernsberger, P., Irwin, J.J. & Shoichet, B.K. (2007). Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, **25**, 197 – 206. 202, 214

Khalifa, A.A., Haranczyk, M. & Holliday, J. (2009). Comparison of nonbinary similarity coefficients for similarity searching, clustering and compound selection. *J. Chem. Inf. Model.*, **49**, 1193–1201, pMID: 19405526. 244

King, A. (2011). Breaking through the barrier. *Chem. World*, **8**, 36–39. 237

Kirew, D.B., Chretien, J.R., Bernard, P. & Ros, F. (1998). Application of kohonen neural networks in classification of biologically active compounds. *SAR QSAR Environ. Res.*, **8**, 93–107, doi: 10.1080/10629369808033262. 47

Kirkpatrick, S., Vecchi, M. *et al.* (1983). Optimization by simmulated annealing. *Science*, **220**, 671–680. 48

Klopman, G. & Zhu, H. (2001). Estimation of the aqueous solubility of organic molecules by the group contribution approach. *J. Chem. Inf. Comput. Sci.*, **41**, 439–445. 226

Kobler, J., Schconing, U. & Toran, J. (1993). *The Graph Isomorphism Problem: Its Structural Complexity*. Progress in Theoretical Computer Science Series, Birkhauser, Boston. 121

Kovdienko, N.A., Polishchuk, P.G., Muratova, E.N., Artemenko, A.G., Kuz'min, V.E., Gorb, L., Hill, F. & Leszczynski, J. (2010). Application of random forest and multiple linear regression techniques to qspr prediction of an aqueous solubility for military compounds. *Mol. Inform.*, **29**, 394–406. 51

Kubiny, H. (1994). Variable selection in qsar studies. i. an evolutionary algorithm. *Mol. Inform.*, **13**, 285–294. 65

Kubinyi, H. (1997a). Qsar and 3d qsar in drug design part 1: methodology. *Drug Discov. Today*, **2**, 457–467. 21

Kubinyi, H. (1997b). Qsar and 3d qsar in drug design part 2: applications and problems. *Drug Discov. Today*, **2**, 538–546. 21

Kubinyi, H. (1998). Similarity and dissimilarity: A medicinal chemist's view. *Perspect. Drug Discov.*, **9-11**, 225–252. 114, 115, 119

Kuhn, H.W. (1955). The hungarian method for the assignment problem. *Nav. Res. Log.*, **2**, 83–97. 126, 127

Kuhne, R., Ebert, R.U., Kleint, F., Schmidt, G. & Schuurmann, G. (1995). Group contribution methods to estimate water solubility of organic chemicals. *Chemosphere*, **30**, 2061–2077. 226

Kuzmin, V., Artemenko, A., Polischuk, P., Muratov, E., Hromov, A., Liahovskiy, A., Andronati, S. & Makan, S. (2005). Hierarchic system of qsar models (1d-4d) on the base of simplex representation of molecular structure. *J. Mol. Model.*, **11**, 457–467. 242

Laarhoven, P.J.M. & Aarts, E.H.L. (1987). *Simulated annealing: theory and applications*. D. Reidel. 48, 49

Lagunin, A., Stepanchikova, A., Filimonov, D. & Poroikov, V. (2000). Pass: prediction of activity spectra for biologically active substances. *Biostatistics*, **16**, 747–748. 23

Laidler, K. (1956). A system of molecular thermochemistry for organic gases and liquids. *Can. J. Chem.*, **34**, 626–648. 243

Lajiness, M.S., Maggiora, G.M. & Shanmugasundaram, V. (2004). Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *J. Med. Chem.*, **47**, 4891–4896. 40, 141, 154, 155

Leach, A.R. & Gillet, V.J. (2007). *An Introduction to Chemoinformatics*. Springer. 25, 26, 31, 32, 33, 36

Leal, J.P. (2006). Additive methods for prediction of thermochemical properties. the laidler method revisited. 1. hydrocarbons. *J. Phys. Chem. Ref. data*, **35**, 55–76. 243

Leardi, R. (2001). Genetic algorithms in chemometrics and chemistry: a review. *J. Chemom.*, **15**, 559–569. 48, 79

Leardi, R. & Lupianez Gonzalez, A. (1998). Genetic algorithms applied to feature selection in pls regression: how and when to use them. *Chemom. Intell. Lab. Syst.*, **41**, 195–207. 48

Li, C. & Colosi, L.M. (2012). Molecular similarity analysis as tool to prioritize research among emerging contaminants in the environment. *Sep. Purif. Technol.*, **84**, 22–28. 155

Li, H., Yap, C.W., Ung, C.Y., Xue, Y., Cao, Z.W. & Chen, Y.Z. (2005). Effect of selection of molecular descriptors on the prediction of blood-brain barrier penetrating and nonpenetrating agents by statistical learning methods. *J. Chem. Inf. Model.*, **45**, 1376–1384. 238

Liaw, A. & Wiener, M. (2002). Classification and regression by randomforest. *R News*, **2**. 71

Lide, D. (2010). *CRC Handbook of Chemistry and Physics*. Boca Raton, FL: CRC Press/Taylor and Francis, 90th edn., (CD-ROM Version). 222

Lill, M.A. (2007). Multi-dimensional qsar in drug discovery. *Drug Discov. Today*, **12**, 1013–1017. 21

Lin, H.Y., Muller, Y.A. & Hammond, G.L. (2010). Molecular and structural basis of steroid hormone binding and release from corticosteroid-binding globulin. *Mol. Cell. Endocrinol.*, **316**, 3 – 12. 234

Linstrom, P. & Mallard, W. ((accessed in 2012)). Nist chemistry webbook, nist standard reference database number 69; national institute of standards and technology, gaithersburg md, 20899. http://webbook.nist.gov/chemistry/. 222

Lipinski, C.A. (2000). Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods*, **44**, 235–249. 226

Lipinski, C.A., Lombardo, F., Dominy, B.W. & Feeney, P.J. (2012). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, **64**, 4–17. 21

Liu, P. & Long, W. (2009). Current mathematical methods used in qsar/qspr studies. *Int. J. Mol. Sci.*, **10**, 1978–1998. 64

Liu, R. & So, S.S. (2001). Development of quantitative structure-property relationship models for early adme evaluation in drug discovery. 1. aqueous solubility. *J. Chem. Inf. Comput. Sci.*, **41**, 1633–1639. 177, 226

Liu, Y. (2004). A comparative study on feature selection methods for drug discovery. *J. Chem. Inf. Comput. Sci.*, **44**, 1823–1828. 63

Maldonado, A.G., Doucet, J.P., Petitjean, M. & Fan, B.T. (2006). Molecular similarity and diversity in chemoinformatics: From theory to applications. *Mol. Divers.*, **10**, 39–79, mol Divers. 154, 155

Maltarollo, V.G., Honorio, K.M. & Silva, A.B.F.d. (2013). Applications of artificial neural networks in chemical problems. In K. Suzuki, ed., *Artificial Neural Networks - Architectures and Applications*, InTech. 47

Martin, Y.C. (1998). 3d qsar: current state, scope, and limitations. In *3D QSAR in Drug Design*, 3–23, Springer. 21

Martin, Y.C., Kofron, J.L. & Traphagen, L.M. (2002). Do structurally similar molecules have similar biological activity? *J. Med. Chem.*, **45**, 4350–4358, doi: 10.1021/jm020155c. 115, 119, 134, 186, 236, 237

Martins, I.F., Teixeira, A.L., Pinheiro, L. & Falcao, A.O. (2012). A bayesian approach to in silico blood-brain barrier penetration modeling. *J. Chem. Inf. Model.*, **52**, 1686–1697. 92, 101, 213, 219, 238

Masterton, W.L. & Hurley, C.N. (2008). *Chemistry: Principles and Reactions*. Cengage Learning. 220

Matheron, G. (1965). *Les Variables Regionalisees et Leur Estimation*. Masson et Cie, Paris. 156, 159

Matter, H. (1997). Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.*, **40**, 1219–1229. 134

Matter, H. & Patter, T. (1999). Comparing 3d pharmacophore triplets and 2d fingerprints for selecting diverse compound subsets. *J. Chem. Inf. Comput. Sci.*, **39**, 1211–1225, doi: 10.1021/ci980185h. 34

McGregor, M.J. & Pallai, P.V. (1997). Clustering of large databases of compounds: Using the mdl "keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.*, **37**, 443–448, doi: 10.1021/ci960151e. 38, 39

Mendeleev, D.I. (1869). The relation between the properties and atomic weights of the elements. *J. Russ. Chem. Soc.*, **1**, 60–77. 40

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. & Leisch, F. (2012). *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. http://cran.r-project.org/web/packages/e1071/e1071.pdf. 70

Meyer, H. (1899). Zur theorie der alkoholnarkose. *N. S. Arch. Pharmacol.*, **42**, 109–118. 20

Micheli, A. (2003). *Recursive processing of structured domains in machine learning*. Ph.D. thesis, University of Pisa. 4

Mills, E.J. (1884). On melting-point and boiling-point as related to chemical composition. *Phil. Mag.*, **17**, 173–187. 20

Mirmehrabi, M., Rohani, S. & Perry, L. (2006). Thermodynamic modeling of activity coefficient and prediction of solubility: Part 2. semipredictive or semiempirical models. *J. Pharm. Sci.*, **95**, 798–809. 226

Mislow, K. & Siegel, J. (1984). Stereoisomerism and local chirality. *J. Am. Chem. Soc.*, **106**, 3319–3328. 138

Mittal, R.R., Harris, L., McKinnon, R.A. & Sorich, M.J. (2009). Partial charge calculation method affects comfa qsar prediction accuracy. *J. Chem. Inf. Model.*, **49**, 704–709. 183

Mosier, P. & Jurs, P. (2002). Qsar/qspr studies using probabilistic neural networks and generalized regression neural networks. *J. Chem. Inf. Comput. Sci.*, **42**, 1460–1470. 64

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *J. Soc. Ind. Appl. Math.*, **5**, 32–38. 126, 127

Negreiros, J., Painho, M., Aguilar, F. & Aguilar, M. (2010). Geographical information systems principles of ordinary kriging interpolator. *J. Appl. Sci.*, **10**, 852–867. 159

Netzeva, T.I., Worth, A.P., Aldenberg, T., Benigni, R., Cronin, M.T., Gramatica, P., Jaworska, J.S., Kahn, S., Klopman, G., Marchant, C.A. *et al.* (2005). Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. In *The report and recommendations of ECVAM Workshop 52*, vol. 33, 155–173, European Union Reference Laboratory for Alternatives to Animal Testing. 59

Niculescu, S.P. (2003). Artificial neural networks and genetic algorithms in qsar. *J. Mol. Struct.*, **622**, 71–83. 47

Nikolova, N. & Jaworska, J. (2003). Approaches to measure chemical similarity - a review. *QSAR Comb. Sci.*, **22**, 1006–1026. 33, 39, 40, 41, 43, 114, 115, 118, 154, 155, 167

NCI/CADD CIR (2011). Chemical identifier resolver beta 4. http://cactus.nci.nih.gov/chemical/structure. 100, 148, 238

O'Boyle, N., Banck, M., James, C., Morley, C., Vandermeersch, T. & Hutchison, G. (2011). Open babel: An open chemical toolbox. *J. Cheminf.*, **3**, Article 33, http://www.jcheminf.com/content/3/1/33 (accessed March, 2013). 29, 37, 100, 135, 136, 149, 168, 241

O'Boyle, N.M., Morley, C. & Hutchison, G.R. (2008). Pybel: a python wrapper for the openbabel cheminformatics toolkit. *J. Cheminf.*, **2**. 29, 100, 136, 241

Obrezanova, O., Csanyi, G., Gola, J.M.R. & Segall, M.D. (2007). Gaussian processes: A method for automatic qsar modeling of adme properties. *J. Chem. Inf. Model.*, **47**, 1847–1857. 156

Oprea, T.I. & Gottfries, J. (2001). Chemography: The art of navigating in chemical space. *J. Comb. Chem.*, **3**, 157–166. 153, 155

Overton, C.E. (1901). *Studien über die Narkose zugleich ein Beitrag zur allgemeinen Pharmakologie*. Fischer. 20

Ozdemir, M., Embrechts, M., Arciniegas, F., Breneman, C., Lockwood, L. & Bennett, K. (2001). Feature selection for in-silico drug design using genetic algorithms and neural networks. In *Proceedings of the 2001 IEEE Mountain Workshop on Soft Computing in Industrial Applications (SMCia/01)*, 53–57, Blacksburg, VA: IEEE, New York City, NY. 48

Palmer, D.S., Llinas, A., Morao, I., Day, G.M., Goodman, J.M., Glen, R.C. & Mitchell, J.B.O. (2008). Predicting intrinsic aqueous solubility by a thermodynamic cycle. *Mol. Pharm.*, **5**, 266–279. 226

Pardridge, W.M., ed. (1998). *Introduction to the Blood-Brain Barrier: Methodology, biology and pathology*. Cambridge University Press. 237, 239

Pardridge, W.M. (2005). The blood-brain barrier: Bottleneck in brain drug development. *NeuroRx*, **2**, 3–14. 63, 237

Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D. & Weinberger, L.E. (1996). Neighborhood behavior: A useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.*, **39**, 3049–3059, doi: 10.1021/jm960290n. 114, 155

Perkins, R., Fang, H., Tong, W. & Welsh, W.J. (2003). Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology. *Environmental Toxicology and Chemistry*, **22**, 1666–1679. 21

Peterangelo, S. & Seybold, P. (2004). Synergistic interactions among qsar descriptors. *Int. J. Quantum Chem.*, **96**, 1–9. 82

Petrone, P.M., Simms, B., Nigsch, F., Lounkine, E., Kutchukian, P., Cornett, A., Deng, Z., Davies, J.W., Jenkins, J.L. & Glick, M. (2012). Rethinking molecular similarity: Comparing compounds on the basis of biological activity. *ACS Chem. Biol.*, **7**, 1399–1409. 202, 214

Prelog, V. & Helmchen, G. (1982). Basic principles of the cip-system and proposals for a revision. *Angew. Chem. Int. Ed. (English)*, **21**, 567–583. 136

Puzyn, T., Leszczynski, J. & Cronin, M. (2009). *Recent Advances in QSAR Studies: Methods and Applications*. London: Springer. 54, 55, 63, 67

R Panico, W.H.P. & Richer, J.C., eds. (1993). *A Guide to IUPAC Nomenclature of Organic Compounds, Recommendations*. Blackwell Scientific Publications. 138

Rahman, S., Bashton, M., Holliday, G., Schrader, R. & Thornton, J. (2009). Small molecule subgraph detector (smsd) toolkit. *J. Cheminf.*, **1**, 12. 118

Rarey, M. & Dixon, J.S. (1998). Feature trees: A new molecular similarity measure based on tree matching. *J. Comput.-Aided Mol. Des.*, **12**, 471–490. 118

Raymond, J. & Willett, P. (2002). Maximum common subgraph isomorphism algorithms for the matching of chemical structures,. *J. Comput.-Aided Mol. Des.*, **16**, 521 – 533. 118

Rennen, G. (2009). Subset selection from large datasets for kriging modeling. *Struct. Multidiscp. Optim.*, **38**, 545–569. 185

Ribeiro Jr, P.J. & Diggle, P.J. (2001). geoR: a package for geostatistical analysis. *R-NEWS*, **1**, 14–18. 166

Robert, D., Amat, L. & Carbo-Dorca, R. (1999). Three-dimensional quantitative structure-activity relationships from tuned molecular quantum similarity measures: prediction of the corticosteroid-binding globulin binding affinity for a steroid family. *J. Chem. Inf. Comput. Sci.*, **39**, 333–344. 234

Rossini, F.D. (1956). *Experimental Thermochemistry*. Interscience Publishers, New York. 220, 221

Rouvray, D.H. (1992). Definition and role of similarity concepts in the chemical and physical sciences. *J. Chem. Inf. Comput. Sci.*, **32**, 580–586, doi: 10.1021/ci00010a002. 39

Roy, P. & Roy, K. (2008). On some aspects of variable selection for partial least squares regression models. *QSAR Comb. Sci.*, **27**, 302–313. 64

Sadowski, J., Gasteiger, J. & Klebe, G. (1994). Comparison of automatic three-dimensional model builders using 639 x-ray structures. *J. Chem. Inf. Comput. Sci.*, **34**, 1000–1008. 168, 240

Salahinejad, M., Le, T.C. & Winkler, D.A. (2013). Aqueous solubility prediction: Do crystal lattice interactions help? *Mol. Pharm.*, **10**, 2757–2766. 226

Santos, R.C., Leal, J.P. & Martinho Simões, J.A. (2009). Additivity methods for prediction of thermochemical properties. the laidler method revisited. 2. hydrocarbons including substituted cyclic compounds. *J. Chem. Thermodyn.*, **41**, 1356–1373. 13, 243

Sharma, M.C., Sharma, S. & Bhadoriya, K.S. (2012). Qsar analyses and pharmacophore studies of tetrazole and sulfonamide analogs of imidazo[4,5-b]pyridine using simulated annealing based feature selection. *J. Saudi Chem. Soc.*. 49

Sheridan, R.P. & Kearsley, S.K. (2002). Why do we need so many chemical similarity search methods? *Drug Discov. Today*, **7**, 903–911. 114, 115

Sorai, M., Atake, T., Inaba, A., Saito, K., Hashimoto, T., Kidokoro, S., Oguni, M., Ozao, R., Tsuji, T., Yokokawa, H. & Yoshida, H. (2004). *Comprehensive Handbook of Calorimetry and Thermal Analysis*. John Wiley & Sons Ltd., Chichester, UK. 220

Soto, A., Cecchini, R., Vazquez, G. & Ponzoni, I. (2009). Multi-objective feature selection in qsar using a machine learning approach. *QSAR Comb. Sci.*, **28**, 1509–1523. 65

Spiess, A.N. & Neumeyer, N. (2010). An evaluation of r2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a monte carlo approach. *BMC Pharmacol.*, **10**, 6. 74

Spitz, E. (1994). *Museums of the Mind: Magritte's Labyrinth and Other Essays in the Arts*. Yale Univerysity Press. 25

Statnikov, A. & Aliferis, C. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, **9**, 319. 54, 65

Stein, S.E., Heller, S.R. & Tchekhovski, D. (2003). An open standard for chemical structure representation - the iupac chemical identifier. In *Nimes International Chemical Information Conference Proceedings*, 131–143. 29

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E. & Willighagen, E. (2003). The chemistry development kit (cdk): An open-source java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.*, **43**, 493–500, pMID: 12653513. 241

Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T. & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, **9**, 307. 71, 72

Stull, D., Westrum, E. & Sinke, G. (1969). *The Chemical Thermodynamics of Organic Compounds*. John Wiley & Sons Inc, New York. 220

Stumpfe, D. & Bajorath, J. (2012). Exploring activity cliffs in medicinal chemistry: Miniperspective. *J. Med. Chem.*, **55**, 2932–2942. 22

Stuper, A.J. & Jurs, P.C. (1976). Adapt: A computer system for automated data analysis using pattern recognition techniques. *J. Chem. Inf. Comput. Sci.*, **16**, 99–105, doi: 10.1021/ci60006a014. 23

Sun, J., Crowe, M. & Fyfe, C. (2010). Extending metric multidimensional scaling with bregman divergences. In *Proceedings of the 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent*

Systems - Volume Part II, IEA/AIE'10, 615–626, Springer-Verlag, Berlin, Heidelberg. 46

Sun, Y., Brown, M., Prapopoulou, M., Davey, N., Adams, R. & Moss, G. (2011). The application of stochastic machine learning methods in the prediction of skin penetration. *Appl. Soft Comput.*, **11**, 2367–2375. 156

Sutherland, J.J., OBrien, L.A. & Weaver, D.F. (2003). Spline-fitting with a genetic algorithm: A method for developing classification structure-activity relationships. *J. Chem. Inf. Comput. Sci.*, **43**, 1906–1915. 182

Sutherland, J.J., OBrien, L.A. & Weaver, D.F. (2004). A comparison of methods for modeling quantitative structure-activity relationships. *J. Med. Chem.*, **47**, 5541–5554. 182, 183, 188, 219, 228

Sutter, J., Dixon, S. & Jurs, P. (1995). Automated descriptor selection for quantitative structure-activity relationships using generalized simulated annealing. *J. Chem. Inf. Comput. Sci.*, **35**, 77–84. 65

Taft Jr, R.W. (1952). Linear free energy relationships from rates of esterification and hydrolysis of aliphatic and ortho-substituted benzoate esters. *J. Am. Chem. Soc.*, **74**, 2729–2732. 21

Takahashi, Y., Sukekawa, M. & Sasaki, S. (1992). Automatic identification of molecular similarity using reduced-graph representation of chemical structure. *J. Chem. Inf. Comput. Sci.*, **32**, 639–643, doi: 10.1021/ci00010a009. 118

Tay, F. & Cao, L. (2001). A comparative study of saliency analysis and genetic algorithm for feature selection in support vector machines. *Intell. Data Anal.*, **5**, 191–209. 48

Teixeira, A.L. & Falcao, A.O. (2013). Noncontiguous atom matching structural similarity function. *J. Chem. Inf. Model.*, **53**, 2511–2524. 119, 155, 169, 213

Teixeira, A.L. & Falcao, A.O. (2014). Structural similarity based kriging for quantitative structure activity and property relationship modeling. *J. Chem. Inf. Model.*, **54**, 1833–1849. 213

Teixeira, A.L., Leal, J.P. & Falcao, A.O. (2013a). Automated identification and classification of stereochemistry: Chirality and double bond stereoisomerism. Tech. rep., Department of Informatics, Faculty of Sciences, University of Lisbon, arXiv:1303.1724. 120, 123, 136, 139, 148, 169, 213

Teixeira, A.L., Leal, J.P. & Falcao, A.O. (2013b). Random forests for feature selection in qspr models - an application for predicting standard enthalpy of formation of hydrocarbons. *J. Cheminf.*, **5**, Article 9, http://www.jcheminf.com/content/5/1/9 (accessed September, 2013). 213

Teixeira, A.L., Santos, R.C., Simoes, J.A.M., Leal, J.P. & Falcao, A.O. (2013c). Therminfo: Collecting, retrieving, and estimating reliable thermochemical data. Tech. rep., Department of Informatics, Faculty of Sciences, University of Lisbon, arXiv:1302.0710. 213, 219, 221, 222, 224

Tetko, I.V., Tanchuk, V.Y., Kasheva, T.N. & Villa, A.E.P. (2001). Estimation of aqueous solubility of chemical compounds using e-state indices. *J. Chem. Inf. Comput. Sci.*, **41**, 1488–1493. 177, 226

Tetko, I.V., Gasteiger, J., Todeschini, R., Mauri, A., Livingstone, D., Ertl, P., Palyulin, V.A., Radchenko, E.V., Zefirov, N.S., Makarenko, A.S., Tanchuk, V.Y. & Prokopenko, V. (2005). Virtual computational chemistry laboratory - design and description. *J. Comput.-Aided Mol. Des.*, **19**, 453–63. 168, 240

Tobler, W. (1970). A computer movie simulating urban growth in the detroit region. *Econ. Geogr.*, **46**, 234–240. 156

Todeschini, R. & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics*. Wiley-VCH, Weinheim. 30, 32, 33, 34, 115, 239, 240

Todeschini, R., Consonni, V., Mannhold, R., Kubinyi, H. & Timmerman, H. (2008). *Handbook of Molecular Descriptors*. Weinheim, Germany: Wiley-VCH Verlag GmbH. 63, 103, 239

Topliss, J.G. & Edwards, R.P. (1979). Chance factors in studies of quantitative structure-activity relationships. *J. Med. Chem.*, **22**, 1238–1244, doi: 10.1021/jm00196a017. 54

Totrov, M. (2008). Atomic property fields: Generalized 3d pharmacophoric potential for automated ligand superposition, pharmacophore elucidation and 3d qsar. *Chem. Biol. Drug. Des.*, **71**, 15–27. 183

Tovar, A., Eckert, H. & Bajorath, J. (2007). Comparison of 2d fingerprint methods for multiple-template similarity searching on compound activity classes of increasing structural diversity. *ChemMedChem*, **2**, 208–217. 116

Tropsha, A. (2006). Variable selection qsar modeling, model validation, and virtual screening. In *Annual reports in computational chemistry*, vol. 2, 113–126, Elsevier: Amsterdam. 54

Tropsha, A. (2010). Best practices for qsar model development, validation, and exploitation. *Mol. Inform.*, **29**, 476–488. 22, 55, 58, 59, 63, 67, 82, 87, 170

Tropsha, A. & Golbraikh, A. (2007). Predictive qsar modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des.*, **13**, 3494–504. 20, 63, 67

Tropsha, A., Gramatica, P. & Gombar, V. (2003). The importance of being earnest: Validation is the absolute essential for successful application and interpretation of qspr models. *QSAR Comb. Sci.*, **22**, 69–77. 54

Truchon, J.F. & Bayly, C.I. (2007). Evaluating virtual screening methods: good and bad metrics for the early recognition problem. *J. Chem. Inf. Model.*, **47**, 488–508. 22

Tuppurainen, K., Viisas, M., Laatikainen, R. & Perakyla, M. (2002). Evaluation of a novel electronic eigenvalue (eeva) molecular descriptor for qsar/qspr studies: Validation using a benchmark steroid data set. *J. Chem. Inf. Comput. Sci.*, **42**, 607–613. 234

Venables, W. & Ripley, B. (2002). *Modern Applied Statistics with S*. Springer, New York. 166

VCCLAB (accessed in 2011). Virtual computational chemistry laboratory. http://www.vcclab.org. 168, 240

Vyas, R., Tambe, S. & Kulkarni, B. (2014). Applications of support vector machines as a robust tool in high throughput virtual screening. *Int. J. of Comp. Biol.*, **1**, 43–55. 70

Wagener, M., Sadowski, J. & Gasteiger, J. (1995). Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic ah receptor activity by neural networks. *J. Am. Chem. Soc.*, **117**, 7769–7775. 234

Walker, J., Carlsen, L. & Jaworska, J. (2003). Improving opportunities for regulatory acceptance of qsars: The importance of model domain, uncertainty, validity and predictability. *QSAR Comb. Sci.*, **22**, 346–350. 153, 155

Walker, T., Grulke, C.M., Pozefsky, D. & Tropsha, A. (2010). Chembench: a cheminformatics workbench. *Biostatistics*, **26**, 3000–3001. 24

Wang, T. & Zhou, J. (1997). Emcss: A new method for maximal common substructure search. *J. Chem. Inf. Comput. Sci.*, **37**, 828–834, doi: 10.1021/ci9601675. 118

Weininger, D. (1988). Smiles, a chemical language and information system .1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–36. 28

Weininger, D., Weininger, A. & Weininger, J.L. (1989). Smiles 2. algorithm for generation of unique smiles notation. *J. Chem. Inf. Comput. Sci.*, **29**, 97–101. 28

Wild, D.J. & Blankey, C.J. (1999). Comparison of 2d fingerprint types and hierarchy level selection methods for structural grouping using ward's clustering. *J. Chem. Inf. Comput. Sci.*, **40**, 155–162. 38

Willett, P. (2004). Evaluation of molecular similarity and molecular diversity methods using biological activity data. In J. Bajorath, ed., *Chemoinformatics : Concepts, Methods, and Tools for Drug Discovery*, vol. 275 of *Methods in Molecular Biology*, 51–63, Humana Press, Totowa, NJ. 134

Willett, P. (2005). Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem.*, **48**, 4183–4199, doi: 10.1021/jm0582165. 114

Willett, P., Barnard, J. & Downs, G. (1998). Chemical similarity searching. *J. Chem. Inf. Comput. Sci.*, **38**, 983 – 996. 118, 134

Wiswesser, W.J. (1954). *A line-formula chemical notation*. Crowell. 27

Wold, S., Sjostrom, M. & Eriksson, L. (2001). Pls-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.*, **58**, 109–130, doi: 10.1016/S0169-7439(01)00155-1. 51

Xu, L. & Zhang, W. (2001). Comparison of different methods for variable selection. *Anal. Chim. Acta*, **446**, 475–481. 65

Xue, L., Godden, J., Gao, H. & Jr, B. (1999). Identification of a preferred set of molecular descriptors for compound classification based on principal component analysis. *J. Chem. Inf. Comput. Sci.*, **39**, 699–704. 64

Yalkowsky, S.H. (1999). *Solubility and Solubilization in Aqueous Media*. Oxford University Press, New York. 226

Yalkowsky, S.H. & Banerjee, S. (1992). *Aqueous solubility: Methods of estimation for organic compounds*. Marcel Dekker, New York. 227

Yamamoto, J.K. (2000). An alternative measure of the reliability of ordinary kriging estimates. *Mathematical Geology*, **32**, 489–509. 165, 166

Yan, A. & Gasteiger, J. (2003). Prediction of aqueous solubility of organic compounds based on a 3d structure representation. *J. Chem. Inf. Comput. Sci.*, **43**, 429–434. 177, 226

Yasri, A. & Hartsough, D. (2001). Toward an optimal procedure for variable selection and qsar model building. *J. Chem. Inf. Comput. Sci.*, **41**, 1218–1227. 63, 68

Young, D.C. (2009). *Computational drug design: a guide for computational and medicinal chemists*. John Wiley & Sons. 61

Zhang, L., Zhu, H., Oprea, T., Golbraikh, A. & Tropsha, A. (2008). Qsar modeling of the blood-brain barrier permeability for diverse organic compounds. *Pharmaceut. Res.*, **25**, 1902–1914. 237, 238

Zhao, Y.H., Abraham, M.H., Ibrahim, A., Fish, P.V., Cole, S., Lewis, M.L., de Groot, M.J. & Reynolds, D.P. (2007). Predicting penetration across the blood-brain barrier from simple descriptors and fragmentation schemes. *J. Chem. Inf. Model.*, **47**, 170–175. 237, 238